ISSN 1592-7415 **eISSN 2282-8214**

Volume n. 56 – 2026

RATIO MATHEMATICA

Journal of Mathematics, Statistics, and Applications

Chief Editors

Šarká Hošková-Mayerová Fabrizio Maturo

Honorary Editor

Franco Eugeni

Webmaster: Fabio Manuppella
Managing Editor: Annamaria Porreca
Legal Manager: Bruna Di Domenico

Publisher

A.P.A.V.
Accademia Piceno – Aprutina dei Velati in Teramo, Italy

RATIO MATHEMATICA
Journal of Mathematics, Statistics, and Applications

ISSN 1592-7415 - eISSN 2282-8214

Ratio Mathematica is a historical journal of Mathematics and Statistics that has been publishing scientific articles for more than thirty years. Ratio Mathematica, founded by Prof. Franco Eugeni, was born in 1990 in Pescara, a city located in the Abruzzo region of Italy. Since the first years of activity, the journal has always been characterized by a strong internationalization.

Since 2013, the journal has been managed by Prof. Fabrizio Maturo, who is a Full Professor in Statistics at the Universitas Mercatorum of Rome (Italy), and Prof. Sarka Hoskova Mayerova, who is a Full Professor in Mathematics at the University of Defence of Brno (Czech Republic).

From 2013 to date, Ratio Mathematica has obtained indexing in many databases such as *DOAJ, PKP (Public Knowledge Project), OCLC WorldCat, Google Scholar, Publons, ROAD, AcademicKeys, Genamics, JournalTOCs, ANCP, ISSN PORTAL, Wikidata, OPAC SBN, Paperity, Mir@bel, zdb-katalog.de, Journal Searches, Proquest, Pubdb, Acnp, Electronic and Magazine Library*.

Ratio Mathematica is included in the list of scientific journals recognized by ANVUR (Italian National Agency of Evaluation of the University and Research system) in:
-Area 13: Economic and statistical sciences.
-Area 11: Historical, philosophical, pedagogical and psychological sciences.

Ratio Mathematica is an open-access journal that publishes an issue every six months (in June and December). The journal is interested in original theoretical and applied Mathematics and Statistics research articles. Only English-language publications are accepted.

The main topics of interest for Ratio Mathematica are:
-Advances in theoretical and applied mathematics and statistics;
-Decision-making in conditions of uncertainty;
-Fuzzy logic;
-Probability theory;
-New theories for the dissemination and communication of mathematics and statistics.

The journal is interested in methodology and applications, but in both cases, a component of originality and innovation is essential to consider the

submission for publication. For this reason, simple applications of known methods are not of interest to the journal.

Webmaster
Manuppella, Fabio, Pescara, Italy. fabiomanuppella@gmail.com

Legal Manager
Di Domenico, Bruna, Pescara, Italy

Note on Peer-Review
All manuscripts are subjected to a double-blind review process. The reviewers are selected from the editorial board, but they also can be external subjects. The journal's policies are described at:
http://eiris.it/ojs/index.php/ratiomathematica/about/submissions#authorGuidelines

Copyright on any research article published by Ratio Mathematica is retained by the author(s). Authors grant Ratio Mathematica a license to publish the article and identify itself as the original publisher. Authors also grant any third party the right to use the article freely as long as its original authors, citation details and publisher are identified.

Publisher: **APAV** - Accademia Piceno-Aprutina dei Velati in Teramo, Italy
Tax Code 92036140678, **Vat Id. Number** 02184450688
Registered office: Contrada Salice, 28 - 87023 - Diamante (CS)
Operational headquarters: c / o Hotel Cristina - via Pietrarossa, 24 - Diamante (CS)
Publishing office: via Pianacci, 24 - 65015 - Montesilvano (PE)

Chief Editors

Mathematics

Hoskova-Mayerova, Sarka, sarka.mayerova@seznam.cz. Department of Mathematics and Physics, University of Defence, Brno, Czech Republic, Brno, Czech Republic.

Google Scholar Profile - SCOPUS Profile

Statistics

Maturo, Fabrizio, fabrizio.maturo@unimercatorum.it, fabmatu@gmail.com, Head of the Faculty of Technological and Innovation Sciences, Department of Economics, Statistics and Business, Universitas Mercatorum, Rome, Italy, Universitas Mercatorum, Piazza Mattei, 10, 00186 Roma, RM, Italy.

Google Scholar Profile - SCOPUS Profile - Web of Science Profile

Founding Editor

Eugeni, Franco, eugenif3@gmail.com. Department of Communication Science, University of Teramo, Teramo, Italy.

Associate Editors

Ameri, Reza, School of Mathematics, Statistics and Computer Sciences, University of Tehran, Teheran, Iran. ameri@ut.ac.ir

Anatriello, Giuseppina, Department of Architecture, University of Naples Federico II, Naples, Italy. giuseppina.anatriello@unina.it

Angelini, Pierpaolo, Sapienza Università di Roma, Rome, Italy. pierpaolo.angelini@istruzione.it

Beutelspacher, Albrecht, Justus Liebig University in Gießen, Giessen, Germany. Albrecht.Beutelspacher@math.uni-giessen.de

Carbó-Dorca, Ramon, Section of Quantum and Mathematical Chemistry, Universitat de Girona, Gerona, Spain. ramoncarbodorca@gmail.com

Casolaro, Ferdinando, Department of Architecture, University of Naples Federico II, Naples, Italy. ferdinando.casolaro@unina.it

Cavallo, Bice, Department of Architecture, University of Naples Federico II, Naples, Italy. bice.cavallo@unina.it

Ciprian, Ionel Alecu, Gheorghe Zane Institute for Economic and Social Research, Iasi, Romania. aiciprian@yahoo.com

Ciuffreda, Raffaela, Department of Law, Economics, Management and Quantitative Methods (D.E.M.M.), University of Sannio, Benevento, Italy. ciuffreda@unisannio.it

Corsini, Piergiulio, Department of Civil Engineering and Architecture, University of Udine, Udine, Italy. piergiulio.corsini@uniud.it

Cristea, Irina, Centre for Systems and Information Technologies, University of Nova Gorica, Nova Gorica, Slovenia. irina.cristea@ung.si

Cruz Rambaud, Salvador, Department of Economics and Business, Almeria, Spain. scruz@ual.es
Cuconato Simone - Department of Humanities, University of Calabria, Cosenza, Italy - simone.cuconato@unical.it,
Darwish Tfiha, Amir, Tishreen University, Syria, Tishreen, Syria. amirtfiha@yahoo.com
Dass, Bal Khishan, Department of Mathematics, University of Delhi, Delhi, India. dassbk@rediffmail.com
De Sanctis, Angela, Department of Business Administration, University of Chieti-Pescara, Pescara, Italy. angela.desanctis@unich.it
Di Battista, Tonio, Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University of Chieti-Pescara, Chieti, Italy. dibattis@unich.it
Dobriţoiu, Maria A., Department of Mathematics and Computer Science, University of Petroşani, Petroşani, Romania. mariadobritoiu@yahoo.com
Dramalidis, Achilles, Department of Education Sciences in Pre-School Age, Democritus University of Thrace, Alexandroupolis, Greece. adramali@psed.duth.gr
Ejegwa, Paul Augustine, Department of Mathematics, University of Agriculture, Makurdi, Nigeria. ejegwa.augustine@uam.edu.ng
Fattoruso, Gerarda, Department of Law, Economics, Management and Quantitative Methods (D.E.M.M.), University of Sannio, Benevento, Italy. fattoruso@unisannio.it
Fensore, Stefania, Department of Economics, University of Chieti-Pescara, Pescara, Italy. stefania.fensore@unich.it
Frezza, Massimiliano, Department of Methods and Models for Economics, Territory and Finance, Sapienza University, Roma, Italy. massimiliano.frezza@uniroma1.it
Gattone, Stefano Antonio, Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, University of Chieti-Pescara, Chieti, Italy. antonio.gattone@unich.it
Husna, V., Department of Mathematics, Presidency University, Itgalpura, Rajanukunte, Yelahanka, Bangalore-560064, India. husna@presidencyuniversity.in
Ibrahim, Abdul, Department of Mathematics, H. H. The Rajah's College Pudukkotai, Tamilnadu, India. dribra@hhrc.ac.in
Jaiswal, Jai Prakash, Department of Mathematics, Guru Ghasidas Vishwavidyalaya (A Central University), Bilaspur, Pantnagar, India. jaiprakashjaiswal@manit.ac.in
Jánský, Jiří, Department of Mathematics and Physics, University of Defence, Brno, Czech Republic. jiri.jansky@unob.cz

Kacprzyk, Janusz, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. kacprzyk@ibspan.waw.pl
Kumar, Manoj, Institute of Engineering & Technology, Lucknow, India. manojkumar@ietlucknow.ac.in
Kumar Singh, Dhiraj, Department of Mathematics, Zakir Husain Delhi College, University of Delhi, Delhi, India. dksingh@zh.du.ac.in
Kumar Singh, Karunesh, Institute of Engineering & Technology, Lucknow, India. kksingh@ietlucknow.ac.in
Manoj, Kumar, Institute of Engineering & Technology, Lucknow, India. manojkumar@ietlucknow.ac.in
Jain, Manish, Ahir College, Rewari, India. jainmanish26128301@gmail.com
Laudano, Francesco, Liceo Classico Mario Pagano, Campobasso, Italy and University of Molise, Italy. francesco.laudano@unimol.it
Mari, Carlo, Department of Economics, University of Chieti-Pescara, Chieti, Italy. carlo.mari@unich.it
Martino, Roberta, roberta.martino@unimercatorum.it, Department of Economics, Statistics, and Business, Universitas Mercatorum, Piazza Mattei, 10, 00186 Roma, RM, Italy.
Martire, Antonio Luciano, Università Roma Tre, Rome, Italy. antonioluciano.martire@uniroma3.it
Mohamed S., Yahya, Department of Mathematics, Government Arts College, Tiruchirappalli, Tamil Nadu, India. yahya_md@yahoo.com
Rackova, Pavlina, Department of Mathematics and Physics, University of Defence, Brno, Czech Republic. pavlina.rackova@unob.cz
Rajeshwari, S., Department of Mathematics, Bangalore Institute of Technology, Bangalore, India. rajeshwaripreetham@gmail.com
Ram, Madhu, University of Jammu, Jammu, India. madhuram0502@gmail.com
Rahman, Bootan, Mathematics Unit, School of Science and Engineering, University of Kurdistan Hewlêr (UKH), Erbil, Iraq. bootan.rahman@ukh.edu.krd
Riccio, Donato, Machine Learning Engineer, Data Science, University of Campania Luigi Vanvitelli, Caserta, Italy, donato_riccio@icloud.com
Rosenberg, Ivo, Université de Montreal, Montreal, Canada. rosenb@dms.umontreal.ca
Sakonidis, Haralambos, Democritus University of Thrace, Komotiní (DUTH), Alexandroupolis, Greece. xsakonid@eled.duth.gr
Scognamiglio, Salvatore, Department of Management and Quantitative Sciences, University of Naples Parthenope, Naples, Italy. salvatore.scognamiglio@uniparthenope.it

Sessa, Salvatore, Department of Architecture, University of Naples Federico II, Naples, Italy. sessa@unina.it
Simonetti, Biagio, Department of Law, Economics, Management and Quantitative Methods (D.E.M.M.), University of Sannio, Benevento, Italy. simonetti@unisannio.it
Squillante, Massimo, Department of Law, Economics, Management and Quantitative Methods (D.E.M.M.), University of Sannio, Benevento, Italy. prorettore@unisannio.it
Sunday, Fadugba, Department of Mathematics, Ekiti State University, Ado Ekiti, Nigeria. sunday.fadugba@eksu.edu.ng
Padiyar, Surendra Vikram Singh, Department of Mathematics, Govt. Degree College Haldwani City, Nainital, Uttarakhand, India. Surendrapadiyar4791@gmail.com
Syed Ali, Muhammad, Department of Mathematics, Thiruvalluvar University, Vellore, India. syedgru@gmail.com
Tallini, Maria Scafati, Department of Mathematics "Guido Catelnuovo", University of Rome La Sapienza, Rome, Italy. tallini@mat.uniroma1.it
Tallini, Luca, Faculty of Communication Sciences, University of Teramo, Teramo, Italy. ltallini@unite.it
Tofan, Ioan, Faculty of Mathematics, Alexandru I. Cuza University, Iasi, Romania. ioantofan@yahoo.it
Tonin, Simone, Department of Economics and Statistics, University of Udine, Udine, Italy. simone.tonin@uniud.it
Udhayakumar, R., Dept of Mathematics, School of Advanced Science, Vellore Institute of Technology (VIT), Tamil Nadu, India. udhayakumar.r@vit.ac.in
Vagaska, Alena, Technical University of Kosice - Technicka univerzita v Kosiciach, Košice, Slovakia. alena.vagaska@tuke.sk
Ventre, Viviana, Department of Mathematics and Physics, University of Campania Luigi Vanvitelli, Caserta, Italy. viviana.ventre@unicampania.it
Ventre, Aldo Giuseppe Saverio, Department of Architecture and Industrial Design, University of Campania Luigi Vanvitelli, Naples, Italy. aldoventre@yahoo.it
Vichi, Maurizio, Department of Statistical Science, University of Rome La Sapienza, Rome, Italy. maurizio.vichi@uniroma1.it
Vincenzi, Giovanni, Department of Mathematics / DIPMAT, University of Salerno, Salerno, Italy. vincenzi@unisa.it
Vougiouklis, Thomas, Department of Primary Level Education, Democritus University of Thrace, Alexandroupolis, Greece. tvougiou@eled.duth.gr
Yager, Ronald R., Iona College, New York City Metro Area Catholic College, New York, U.S.A. ryager@iona.edu

Zifaro, Maria, Faculty of Society and Communication Sciences, Universitas Mercatorum, Rome, Italy. maria.zifaro@unimercatorum.it

Managing Editor
Porreca, Annamaria, Department of Medical, Oral and Biotechnological Sciences, Laboratory of Biostatistics, University "G. d'Annunzio" Chieti-Pescara, Chieti, Italy. annamaria.porreca@unich.it.

Webmaster
Manuppella, Fabio, Pescara, Italy

Legal Manager
Di Domenico, Bruna, Pescara, Italy
Participation in the editorial board is renewed annually based on the contribution and participation.

Ratio Mathematica Volume 56, 2026

Common Fixed Point Theorems For Ćirić-Reich-Rus Type Contractions In F-Metric Spaces

Canan Acar*
Vildan Ozturk †

Abstract

In this work, we introduce a Ćirić-Reich-Rus type contraction for four mappings and prove a common fixed point theorem for mappings satisfying this contractive condition in F-metric spaces. Additionally, we present several related results. To illustrate the validity and applicability of our theoretical findings, we provide an application that bridges abstract theory and practical implementation.
Keywords: F-metric, common fixed point, weakly compatible mappings.
2020 AMS subject classifications: 47H10, 54H25. [1]

*Department Of Mathematics, Institute of Graduate Programs, Ankara Haci Bayram Veli University, Ankara, Turkey; cananacarr@icloud.com
†Department Of Mathematics, Faculty of Polatlı Science And Letters, Ankara Haci Bayram Veli University, Ankara, Turkey; vildan.ozturk@hbv.edu.tr
[1]Received on January 1, 2026. Accepted on May 15, 2026. Published on June 30, 2026.
DOI: 10.23755/rm.v56i0.1736. ISSN: 1592-7415. eISSN: 2282-8214.
©Acar and Ozturk. This paper is published under the CC-BY licence agreement.

1 Introduction

The notion of F-metric space was introduced as a generalization of metric spaces. F-metric spaces satisfy the properties of non-negativity, identity, symmetry and a generalized triangular inequality by means of an F-function defined by Wardowski. The authors also established the Banach contraction principle in F-metric spaces in 2018 [Jleli and Samet, 2018]. Later Bera et al. were interested in the topological properties of F-metric spaces[Bera et al., 2019].
Let Ω denote the family of all functions $F : (0, \infty) \to \mathbb{R}$ satisfying the following properties;

F_1) F is increasing,

F_2) For every sequence $\{x_n\}_{n\in\mathbb{N}}$, $\lim_{n\to\infty} x_n = 0 \Leftrightarrow \lim_{n\to\infty} F(\ x_n) = -\infty$ [Wardowski, 2012].

Definition 1.1. *[Jleli and Samet, 2018] Let X be a nonempty set and let $(F, \alpha) \in \Omega \times [0, +\infty)$. A mapping $d : X \times X \to [0, \infty)$ is called an F-metric if it satisfies:*

D_1) $d(x, y) = 0 \Leftrightarrow\ x = y$,

D_2) $d(x, y) = d(y, x)$

D_3) *for $\forall n \geq 2$ ($n \in \mathbb{N}$) and $\forall \{u_i\} \subset X$ with $(u_1, u_n) = (x, y)$ such that*

$$d(x, y) > 0 \Longrightarrow F(d(x, y)) \leq F\left(\Sigma_{i=1}^{n-1} d(u_i, u_{i+1})\right) + \alpha$$

for all $x, y \in X$. In this case (X, d) is called an F-metric space.

Definition 1.2. *[Jleli and Samet, 2018] Let $\{x_n\} \subset X$ be a sequence.*

i. *$\{x_n\}$ is said to be F-convergent if there exists a $x \in X$ such that $d(x_n, x) \to 0$ as $n \to \infty$.*

ii. *$\{x_n\}$ is said to be an F-Cauchy sequence if $d(x_n, x_m) \to 0$ as $n, m \to \infty$.*

iii. *(X, d) is said to be F-complete if every F-Cauchy sequence is F-convergent.*

Theorem 1.1. *[Bera et al., 2019] Every $F-$metric space is Hausdorff.*

Theorem 1.2. *[Jleli and Samet, 2018] Let (X, d) be an $F-$complete $F-$metric space and let $T : X \to X$ be a self-mapping. Suppose that*

$$d(Tx, Ty) \leq ad(x, y) \tag{1}$$

for all $x, y \in X$, where $a \in (0, 1)$. Then T has a unique fixed point.

Many researchers have developed various contraction principles and proved fixed point results for a self-mapping in F-metric spaces. In F-metric spaces, Lateefa [Lateefa and Ahmad, 2019] and [Mitrovic et al., 2019] generalized (1) as the following contractions:

i. Das Gupta contraction:

$$d(Tx,Ty) \leq a_1 d(x,y) + a_2 \frac{[1+d(x,Tx)]\, d(y,Ty)}{1+d(x,y)}$$

where $a_1 \in (0,1)$ with $a_1 + a_2 < 1$.

ii. Reich contraction:

$$d(Tx,Ty) \leq a_1 d(x,y) + a_2 d(x,Tx) + a_3 d(y,Ty)$$

where $a_1, a_2, a_3 \in (0,1)$ with $a_1 + a_2 + a_3 < 1$.

Al-Mezel et al. [Mezel et al., 2020], Faraji et al. [Faraji et al., 2022] defined $\alpha-\beta-$admissible-type contraction. Acar and Ozturk [Acar and Ozturk, 2024] defined $\alpha-$admissible Berinde type contraction and Hussain and Kanwal [Hussain and Kanwal, 2018] defined $\alpha-$admissible Ćirić type contraction. Ozturk [Ozturk, 2023] defined Ćirić-Presic type contraction. Mitrovic et al. [Mitrovic et al., 2019] introduced following contraction principle and proved common fixed point theorems in F-metric spaces.

Theorem 1.3. *[Mitrovic et al., 2019] Let (X,d) be an F-complete F-metric space and let $T, S : X \to X$ be self-mappings satisfying*

$$d(Tx,Ty) \leq a_1 d(Sx,Sy) + a_2 d(Sx,Tx) + a_3 d(Sy,Ty)$$

for all $x, y \in X$, where $a_1 \in (0,1)$ and $a_2, a_3 \in [0,1)$ with $a_1 + a_2 + a_3 < 1$.

In this work, we generalize the concept of Ćirić-Reich-Rus type contractions in F-metric spaces for four mappings and we prove a fixed point theorem for weakly compatible mappings.

Definition 1.3. *[Jungck, 1996] Let X be a nonempty set. Two mappings $T, S : X \to X$ are said to be weakly compatible if $TSx = STx$ whenever $Tx = Sx$.*

2 Main Results

Definition 2.1. *Let (X,d) be an F-complete F-metric space. The self-mappings $T, S, A, B : X \to X$ are called a Ćirić-Reich-Rus type contraction if there exist $a_1 \in (0,1)$ and $a_2, a_3 \in [0,1)$ with $a_1 + a_2 + a_3 < 1$ such that*

$$d(Ax,By) \leq a_1 d(Sx,Ty) + a_2 d(Ax,Sx) + a_3 d(By,Ty) \tag{2}$$

for all $x, y \in X$.

Theorem 2.1. *Let (X, d) be an F-complete F-metric space. Let $T, S, A, B : X \to X$ be Ćirić-Reich-Rus type contraction mappings (2). Assume that*

i. $A(X) \subseteq T(X), B(X) \subseteq S(X)$,

ii. $T(X)$ *or* $S(X)$ *is a closed subset of* X,

iii. the pairs (A, S) *and* (B, T) *are weakly compatible.*

Then A, B, S and T have a unique common fixed point.

Proof

Let $x_0 \in X$ be arbitrary. Since $Ax_0 \in T(X)$, there exists $x_1 \in X$ such that $Tx_1 = Ax_0$. Similarly, since $Bx_1 \in S(X)$, there exists $x_2 \in X$ such that $Sx_2 = Bx_1$. Continuing inductively, we obtain a sequence $\{x_n\}$ satisfying $Tx_{2k+1} = Ax_{2k}$ and $Sx_{2k+2} = Bx_{2k+1}$, $k = 0, 1, 2, \ldots$.
Define

$$y_{2k} = Tx_{2k+1} = Ax_{2k}$$

$$y_{2k+1} = Sx_{2k+2} = Bx_{2k+1}.$$

Now we show that $\{y_n\}$ is an F-Cauchy sequence. Let $(F, \alpha) \in \Omega \times [0, \infty)$ be the pair satisfying condition (D_3). By (F_2), there exists $\delta > 0$ such that

$$0 < t < \delta \Rightarrow F(t) < F(\epsilon) - \alpha. \tag{3}$$

Using the contractive condition (2),

$$\begin{aligned} d(y_{2k}, y_{2k+1}) &= d(Ax_{2k}, Bx_{2k+1}) \\ &\leq a_1 d(Sx_{2k}, Tx_{2k+1}) + a_2 d(Ax_{2k}, Sx_{2k}) + a_3 d(Bx_{2k+1}, Tx_{2k+1}) \\ &= a_1 d(y_{2k-1}, y_{2k}) + a_2 d(y_{2k}, y_{2k-1}) + a_3 d(y_{2k+1}, y_{2k}). \end{aligned}$$

Thus, we get

$$(1 - a_3) d(y_{2k}, y_{2k+1}) \leq (a_1 + a_2) d(y_{2k}, y_{2k-1}).$$

So, we have

$$d(y_{2k}, y_{2k+1}) \leq \left(\frac{a_1 + a_2}{1 - a_3} \right) d(y_{2k}, y_{2k-1}).$$

Set $d_{2k} = d(y_{2k}, y_{2k+1})$

$$d_{2k} \leq \left(\frac{a_1 + a_2}{1 - a_3} \right) d_{2k-1}.$$

Since $a_1 + a_2 + a_3 < 1$, we get $d_{2k} \leq d_{2k-1}$ for $a = \frac{a_1+a_2}{1-a_3} < 1$. Similarly, it is easy to see that $d_{2k+1} \leq d_{2k}$. Repeating this process n times we obtain

$$\begin{aligned} d(y_{2k}, y_{2k+1}) &\leq a(y_{2k-1}, y_{2k}) \\ &\leq a^2 d(y_{2k-2}, y_{2k-1}) \\ &\cdot \\ &\cdot \\ &\cdot \\ &\leq a^n d(y_0, y_1). \end{aligned}$$

Hence, for $m > n$

$$\sum_{i=n}^{m-1} d(y_i, y_{i+1}) \leq \left(\frac{a^n}{1-a}\right) d(y_0, y_1).$$

Since $a < 1$,

$$\lim_{n\to\infty} \left(\frac{a^n}{1-a}\right) d(y_0, y_1) = 0.$$

There exists some $N \in \mathbb{N}$ such that

$$0 < \left(\frac{a^n}{1-a}\right) d(y_0, y_1) < \delta$$

for $n \geq N$. By (3) and $(F1)$, for $m > n > N$, we have

$$F\left(\sum_{i=n}^{m-1} d(y_i, y_{i+1})\right) \leq F\left(\left(\frac{a^n}{1-a}\right) d(y_0, y_1)\right) \leq F(\epsilon) - \alpha. \tag{4}$$

Using (D_3) and (4),

$$d(y_n, y_m) > 0 \Rightarrow F\left(d(y_n, y_m)\right) \leq F\left(\sum_{i=n}^{m-1} d(y_i, y_{i+1})\right) + \alpha < F(\epsilon).$$

By $(F1)$, we get $d(y_n, y_m) < \epsilon$.
Thus $\{y_n\}$ is an F-Cauchy sequence. By F-completeness of (X, d) there exists $y \in X$ such that

$$\lim_{k\to\infty} Tx_{2k+1} = \lim_{k\to\infty} Ax_{2k} = \lim_{k\to\infty} Bx_{2k+1} = \lim_{k\to\infty} Sx_{2k+2} = y.$$

Assume without loss of generality that $T(X)$ is a closed subset of X. Then there exists $v \in X$ such that $Tv = y$.
Suppose $Bv \neq y$. Applying the contractive condition (2) we have

$$d(Ax_{2k}, Bv) \leq a_1 d(Sx_{2k}, Tv) + a_2 d(Ax_{2k}, Sx_{2k}) + a_3 d(Bv, Tv).$$

Letting $k \to \infty$ yields

$$d(y, Bv) \leq a_1 d(y, Tv) + a_2 d(y, y) + a_3 d(Bv, Tv)$$

and

$$d(y, Bv) \leq a_1 d(y, y) + a_2 d(y, y) + a_3 d(Bv, y).$$

So

$$d(y, Bv) \leq a_3 d(Bv, y)$$

which is contradiction since $a_3 < 1$. Hence, $Bv = y$. Since (B, T) is weakly compatible, we have $Ty = TBv = BTv = By$.
Now we show that $By = y$. Apply the contractive condition (2)

$$d(Ax_{2k}, By) \leq a_1 d(Sx_{2k}, Ty) + a_2 d(Ax_{2k}, Sx_{2k}) + a_3 d(By, Ty)$$

Letting $k \to \infty$, we get

$$d(y, By) \leq a_1 d(y, Ty) + a_2 d(y, y) + a_3 d(By, Ty).$$

Thus we get

$$d(y, By) \leq 0.$$

Hence $y = By$. Similarly, we show that y is also fixed point of A and S. Therefore $Ay = Ty = Sy = By = y$.
Now, we show that y is unique fixed point. Suppose z is another common fixed point of A, B, S, T. Then

$$d(z, y) = d(Az, By) \leq a_1 d(Sz, Ty) + a_2 d(Az, Sz) + a_3 d(By, Ty) = 0.$$

So $d(z, y) = 0$. Hence $z = y$.

Example 2.1. *Let $X = [0, 1]$ be an F-metric space equipped with the F-metric $d(x, y) = \begin{cases} 2^{|x-y|}, & \text{if } x \neq y, \\ 0, & \text{if } x = y, \end{cases}$ and let $F(t) = \frac{-1}{t}$ with $\alpha = 1$. Define the mappings $A, B, S, T : X \to X$ by*

$$A(x) = \left(\frac{x}{2}\right)^8, \quad B(x) = \left(\frac{x}{2}\right)^2, \quad T(x) = x^2, \text{ and } S(x) = \left(\frac{x}{2}\right)^2.$$

The pairs (A, S) and (B, T) are weakly compatible, and both $T(X)$ and $S(X)$ are closed subsets of X. Furthermore, for $x \neq y$, the contractive condition holds as follows for suitable $a_1, a_2, a_3 > 0$ with $a_1 + a_2 + a_3 < 1$:

$$\begin{aligned} d(Ax, By) &= 2^{\left|\left(\frac{x}{2}\right)^8 - \left(\frac{y}{2}\right)^2\right|} = a_1 2^{\left|\left(\frac{x}{2}\right)^2 - y^2\right|} + a_2 2^{\left|\left(\frac{x}{2}\right)^8 - \left(\frac{x}{2}\right)^2\right|} + a_3 2^{\left|\left(\frac{y}{2}\right)^2 - y^2\right|} \\ &\leq a_1 d(Sx, Ty) + a_2 d(Ax, Sx) + a_3 d(By, Ty). \end{aligned}$$

$x = 0$ is the unique common fixed point of A, S, T, B.

If we take $S = T = I$ identity map in Theorem 2.1, we obtain following corollary.

Corollary 2.1. *Let (X, d) be an F-complete F-metric space and let $A, B : X \to X$ satisfy*

$$d(Ax, By) \leq a_1 d(x, y) + a_2 d(Ax, x) + a_3 d(By, y) \tag{5}$$

where $a_1, a_2, a_3 > 0$ with $a_1 + a_2 + a_3 < 1$. Then A and B have a unique common fixed point.

If we take $A = B$ in Corollary 2.1, we obtain following corollary.

Corollary 2.2. *Let (X, d) be an F-complete F-metric space and let $A : X \to X$ satisfies*

$$d(Ax, Ay) \leq a_1 d(x, y) + a_2 d(Ax, x) + a_3 d(Ay, y) \tag{6}$$

where $a_1, a_2, a_3 > 0$ with $a_1 + a_2 + a_3 < 1$. Then A has a unique fixed point.

Corollary 2.3. *Let (X, d) be an F-complete F-metric space and let $A, B, S, T : X \to X$ satisfy*

$$d(Ax, By) \leq ad(Sx, Ty) \tag{7}$$

where $0 < a < 1$ with the same conditions (i) and (ii) as in Theorem 2.1. Then A, B, S and T have a unique common fixed point.

3 Application

Let us consider the integral equation

$$x(t) = \alpha \int_0^1 T_i(t, s, x(s))\, ds + \beta \int_0^1 W_j(t, s, x(s))\, ds, i, j = 1, 2 \tag{8}$$

where α and β are real numbers and $T_i, W_j : [0, 1] \times [0, 1] \times \mathbb{R} \to \mathbb{R}^+$ is continuous function. Equip $C[0, 1]$ with the F-metric $d(x, y) = |x - y|^2$.

Theorem 3.1. *Suppose there exists $\gamma \in [0, 1)$ such that for all $t, s \in [0, 1]$ and $x, y \in C[0, 1]$,*

$$|T_1(t, s, x(s)) - T_2(t, s, y(s))| \leq \sqrt{\gamma}\, |x(t) - y(t)|,$$

$$|W_1(t, s, x(s)) - W_2(t, s, y(s))| \leq \sqrt{\gamma}\, |x(t) - y(t)|.$$

Then the integral equation (8) has a unique solution with $\left(\frac{\alpha\sqrt{\gamma}}{1-\beta\sqrt{\gamma}}\right)^2 < 1$.

Proof. Define the operators

$$\begin{aligned} Ax(t) &= \alpha \int_0^1 T_1(t,s,x(s))\,ds \\ Bx(t) &= \alpha \int_0^1 T_2(t,s,x(s))\,ds \\ Sx(t) &= Ix(t) - \beta \int_0^1 W_1(t,s,x(s))\,ds \\ Tx(t) &= Ix(t) - \beta \int_0^1 W_2(t,s,x(s))\,ds. \end{aligned}$$

For all $x, y \in X$ and $t \in [0,1]$, we have

$$\begin{aligned} |Ax(t) - By(t)|^2 &= \left| \begin{array}{c} \alpha \int_0^1 T_1(t,s,x(s))\,ds \\ -\alpha \int_0^1 T_2(t,s,y(s))\,ds \end{array} \right|^2 \\ &\leq \alpha^2 \int_0^1 |T_1(t,s,x(s)) - T_2(t,s,y(s))|^2\,ds \\ &\leq \alpha^2\gamma \int_0^1 |x(t)-y(t)|^2\,ds \\ &\leq \alpha^2\gamma\,|x(t)-y(t)|^2 \int_0^1 ds. \end{aligned}$$

Thus, we get

$$d(Ax,By) \leq \alpha^2\gamma d(x,y). \tag{9}$$

Similarly,

$$\begin{aligned} |Sx(t) - Ty(t)|^2 &= \left| \begin{array}{c} Ix(t) - \beta \int_0^1 W_1(t,s,x(s))\,ds \\ -Iy(t) + \beta \int_0^1 W_2(t,s,y(s))\,ds \end{array} \right|^2 \\ &= \left| x(t) - y(t) - \beta \left(\int_0^1 [W_1(t,s,x(s)) - W_2(t,s,y(s))\,ds] \right) \right|^2 \\ &\geq \left[|x(t)-y(t)| - \left| \beta \left(\int_0^1 [W_1(t,s,x(s)) - W_2(t,s,y(s))\,ds] \right) \right| \right]^2 \\ &\geq \left[(1-\beta\sqrt{\gamma})\,|x(t)-y(t)| \right]^2. \end{aligned}$$

Hence we have

$$d(Sx,Ty) \geq (1-\beta\sqrt{\gamma})^2\, d(x,y).$$

From (9),

$$d(Ax,By) \leq \alpha^2\gamma \frac{d(Sx,Ty)}{(1-\beta\sqrt{\gamma})^2} = \left(\frac{\alpha\sqrt{\gamma}}{1-\beta\sqrt{\gamma}} \right)^2 d(Sx,Ty)$$

All conditions of Corollary 2.3 are satisfied. Therefore integral equation (8) has a unique solution.

4 Discussion and Conclusions

Banach's fixed point theorem is a fundamental result guaranteeing the existence and uniqueness of fixed points. Over the last century, this result has been generalized to two, three, and four mappings, yielding various common fixed point theorems. The present study is the first to extend the Ćirić-Reich-Rus type contraction from two mappings to four mappings in the setting of F-metric spaces. An application to an integral equation is provided to demonstrate the practical utility of the theoretical results. The findings obtained here will serve as a guide for future research on fixed point theory in generalized metric spaces and will contribute to the solution of more complex problems in nonlinear analysis and its applications.

Acknowledgements

The authors would like to thank the anonymous referee for the valuable comments that helped improve this article.

References

C. Acar and V. Ozturk. Fixed point theorems for almost alpha-admissible mappings in f-metric spaces. *Fundamental Journal of Mathematics and Applications*, 7(4), 2024.

A. Bera, H. Garai, B. Damjanovic, and A. Chanda. Some interesting results on f-metric spaces. *Filomat*, 33(10):3257–3268, 2019.

H. Faraji, N. Mirkov, Z. Mitrović, R. Ramaswamy, O. Abdelnaby, and S. Radenović. Some new results for (alpha,beta)-admissible mappings in f-metric spaces with applications to integral equations. *Symmetry*, 14(11):2429, 2022.

A. Hussain and T. Kanwal. Existence and uniqueness for a neutral differential problem with unbounded delay via fixed point results. *Trans. A. Razmadze Math. Institute*, 172(3):481–490, 2018.

M. Jleli and B. Samet. On a new generalization of metric spaces. *J. Fixed Point Theory Appl.*, 20:128, 2018.

G. Jungck. Common fixed points for noncontinuous nonself mappings on nonnumeric spaces. *Far East J. Math. Sci.*, 4(2):199–212, 1996.

D. Lateefa and J. Ahmad. Dass and gupta's fixed point theorem in f-metric spaces. *J. Nonlinear Sci. Appl.*, 12:405–411, 2019.

S. Mezel, J. Ahmad, and G. Marino. Fixed point theorems for generalized (alpha-beta-psi)-contractions in f-metric spaces with applications. *Mathematics*, 8(4): 584, 2020.

Z. Mitrovic, H. Aydi, N. Hussain, and A. Mukheimer. Reich, jungck, and berinde common fixed point results on f-metric spaces and an application. *Mathematics*, 7:387, 2019.

V. Ozturk. Some results for Ćirić prešić type contractions in f-metric spaces. *Symmetry*, 15(8):1521, 2023.

D. Wardowski. Fixed points of a new type of contractive mappings in complete metric spaces. *Fixed Point Theory Appl.*, 94, 2012.

Ratio Mathematica Volume 56, 2026

Exploring Markov chain and Random Forest to predict loan default in rural banks

Joseph Kwabina Arhinful Johnson*
Ellis Agyeman†
Samuel Kwame Okai‡

Abstract

This study investigates loan default risks in rural banks in Ghana's by developing a hybrid predictive model that combines Markov chains and Random Forest machine learning. The Markov chain analysis examines the transition of loans between risk categories (Current, OLEM, Sub-standard, Doubtful, and Loss) based on Bank of Ghana (BoG) standards, revealing the dynamic nature of credit risk over time. The Random Forest model then classifies loans as default or non-default, identifying key features, with Markov chain transition probabilities serving as a crucial input. Using a dataset of 25,981 loan records from a rural bank, the study finds that a significant portion of borrowers operate in non-traditional sectors, with 14% of loans classified as Losses. The Markov analysis indicates stability in the Current and Loss states, while other risk categories exhibit notable transitions. The Random Forest classifier, utilizing features like Days in Arrears and Interest Charge Type, achieves 93% accuracy and strong F1 scores (0.945 for non-defaults and 0.865 for defaults). The model's Receiver Operating Characteristic and Area Under the Curve (ROC- AUC) score of 0.90 demonstrates its excellent ability to distinguish between defaulting and non-defaulting loans. The model identifies loan duration and arrears as primary predictors of default. Scenario simulations confirm that default risk increases as the loan ages and arrears accumulate. Overall, the hybrid model offers superior predictive accuracy and granular risk insights, enhancing credit risk

*Takoradi Technical University, Takoradi, Ghana; joseph.arhinful@ttu.edu.gh.
† Takoradi Technical University, Takoradi, Ghana; pk.akroma@gmail.com.
‡Takoradi Technical University, Takoradi, Ghana; oksaisamuel415@gmail.com.

management for rural banks and supporting financial inclusion in underserved communities.

Keywords: Markov chain, Random Forest, Loan Default[§]

1. Introduction

Individuals worldwide rely on bank credit to address short-term liquidity gaps and pursue long-term goals (Vermoesen et al., 2013). In a competitive and volatile financial environment, borrowing has become essential for households and firms, while lending remains a primary driver of bank profitability (Ekpu & Paloni, 2016). Interest income from loans constitutes a significant share of banks' revenues, but it also exposes them to credit risk—the possibility that borrowers will fail to meet contractual repayment obligations (Brown & Moles, 2014).

To mitigate this risk, lenders must rigorously assess applicants before disbursing funds. Traditionally, loan officers have used the "5 Cs" of credit—character, capital, capacity, collateral, and conditions—to gauge borrower quality (Oseni, 2023). However, this judgment-driven approach is limited by human bias and inconsistency, particularly with increasing loan (Abraham, 2023).

Many banks have employed specialized analysts to review each case and assign a numerical credit score that summarizes default risk (Golin & Delhaise, 2013). These scores, derived from variables like past repayment behavior and income stability, serve as probabilistic estimates of a borrower's ability and willingness to repay on time (Skiba & Tobacman, 2008).

Rural banks play a crucial role in promoting financial inclusion, particularly among low-income populations in rural communities. However, they face high loan default rates due to limitations in traditional credit assessment models and the lack of reliable predictive tools (Agyapong, 2023). Previous studies have primarily utilized traditional financial models, which often fail to capture the dynamic and temporal nature of loan defaults.

This study aims to develop and evaluate a hybrid model combining Markov chain and Random Forest techniques to predict loan defaults particularly for rural banks. By considering transitional probabilities and identifying key influencing factors, this research will fill existing gaps and offer a novel approach to credit risk management.

The hybrid framework is used to integrate the Random Forest and Markov chain models, with the computed transition probabilities serving as a crucial input for the

[1] Received on January 01, 2025. Accepted on May 31, 2025. Published on June 30, 2026.
DOI: 10.23755/rm.v56i0.1744. ISSN 1592-7415; e-ISSN 2282-8214.
©Joseph Kwabina Arhinful Johnson et al.. This paper is published under the CC-BY licence agreement.

machine learning classifier. The first step in this method is to model the stochastic movement of loans between five regulatory risk states: Current, OLEM, Sub-Standard, Doubtful, and Loss using a Discrete-Time Markov Chain (DTMC).

In this hybrid architecture, the Random Forest algorithm uses the transition probabilities obtained from the Markov analysis as input features in order to compute a comprehensive default risk score (RS) in conjunction with other borrower-specific covariates such as Days in Arrears and Interest Charge Type. The transition probabilities assist the model account for the dynamic and temporal aspect of credit risk, which standard static models frequently fail to capture. This score is the weighted sum of these features.

The work advances from merely forecasting the next likely state to offering a binary classification of whether or not a loan will eventually fail by including these transition-based insights into the predictive model. This combination improves the accuracy of identifying high-risk accounts before they enter the "Loss" state by enabling the bank to use borrower-level risk scores in conjunction with long-term transition dynamics for daily decision-making.

The hybrid model will address the limitations of traditional credit assessment models and provide a more accurate prediction of loan defaults. By leveraging the strengths of both Markov chains and Random Forest algorithms, this study will contribute to the development of more effective credit risk management strategies for rural banks.

2. Objectives of the study

The primary objective of this study is to develop a Markov Chain model for predicting loan defaults at rural banks. Consequently, the study seeks to achieve the following specific objectives:

1. Develop a Markov chain clustering model to predict loan defaults effectively.
2. Identify key features influencing the loan default prediction using Random Forest.
3. Improve prediction using optimized hyperparameter tuning.

3. Materials and methods

This study employed a quantitative research approach with a descriptive and exploratory design to analyze historical loan repayment data and predict default outcomes using the Markov Chain, K-Means Clustering, and Random Forest model. The design enabled the identification of trends and patterns in clients' loan repayment behavior and facilitated the application of the Markov Chain to model state transitions.

The study was conducted at a rural bank in Ghana's Western Region, serving low-income populations. The bank offers various financial services, including microloans and SME financing. The population comprised 25,981 loan clients with complete repayment histories. The study used a census approach, analyzing data from all clients.

Data processing involved coding, editing, and cleaning to ensure accuracy. Loan repayment behavior was categorized into five states: Current, OLEM, Sub-Standard, Doubtful, and Loss/Default. The Markov chain model estimated transition probabilities, and steady-state probabilities assessed long-term behavior patterns.

Analytical tools included Python and Microsoft Excel. The study addressed research objectives through probabilistic modeling, with justifications based on theoretical and empirical literature. This approach provided a structured framework for predictive modeling and enhanced reproducibility.

3.1 Discrete-Time Markov Chain (DTMC)

A stochastic process $\{X_t\}t \geq 0,$ where X_t represents the state of the process at time t, is called a DTMC if it satisfies the following conditions:

1. State Space (S): The process takes values from a finite or countable set of loan status. These states are defined as

$$S = \{S_1, S_2, S_3, S_4, S_5\} \tag{1}$$

where the five states are {current, Olem, Sub-standard, Doubtful, and loss} respectively.

2. The Markov property

$$P(S_{n+1=j} \mid S_{n=i}, S_{n-1} = k, \ldots S_{0=X0}) = P(S_{n+1=j} \mid S_{n=i}) \tag{2}$$

for all states i and j being members in Equation (1) $(i, j \in S)$ at all times $(t \geq 0)$.

3.1.1 Transition Probabilities (P)

The matrix (P_{ij}) contains the probabilities of transitioning from state i to j within a one step as indicated in Equation (3).

$$P = [P_{ij}] \tag{3}$$

From Equation (3):

$$\hat{P}_{ij} = \frac{n_{ij}}{\sum_{k \in S} n_{ik}} \tag{4}$$

Where n_{ij} is the observed count or number of transitions from state i to j over the study period, and $\sum_{k \in S} n_{ik}$ is the total number of transitions which originates from state i.

However, P_{ij} satisfies the condition of probability distribution:

$$P_{ij} \geq 0 \tag{5}$$

$$\sum_{j \in S} P_{ij} = 1 \tag{6}$$

for all $i, j \in S$.

3.1.2 Random Forest

Random Forest is an ensemble learning method. It builds many decision trees and combines their predictions to improve accuracy and avoid over fitting.

3.1.2.1 Prediction equations (Ensemble Prediction)

Let x be any input vector containing loan features like days in arrest, total age of loan, etc. The ensemble of B decision trees $\{h_b(x)\}_{b=1}^{B}$ produces a final classification $\hat{y}$ through majority vote given as:

$$\hat{y} = \text{mod}\{h_1(x), h_2(x), ..., h_B(x)\} \tag{7}$$

where $h_b(x)$ is the prediction from the b^{th} decision tree.

For Regression, the prediction is the average of all tree predictions given as

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} h_b(S) \tag{8}$$

3.1.2.2 Prediction Equations (Risk Score - RS)

At this point, a customer's loan risk score (RS) is calculated as a weighted linear combination of its features given as:

$$RS = \sum_{i=1}^{n} \upsilon_i \tau_i \tag{9}$$

where, n is the total number of the influencing features. Example is $n = 10$ for the top ten identified features. Also, υ_i is the importance weight of feature i. An example is the weight of 0.3003 for the number of days in arrears, and τ_i is the normalized value of feature i for a specific borrower.

3.1.2.3 Prediction Equations (Probability of Default - PD)

We further mapped the RS to a probability scale using a logistic (sigmoid) function given as:

$$P(Default) = \frac{1}{1 + e^{-RS}} \tag{10}$$

where, e is the Euler's constant approximately 2.718, and RS is the calculated RS from the Random Forest model in Equation (9).

3.1.2.4 Prediction Equations (Time-Based Risk Evolution)

The time-based risk evolution representing a power-law growth model is presented as:

$$Risk(t) = B_r \times (1 + D_a)^{\alpha} (1 + \beta L_a)^{\beta} \tag{11}$$

where, $Risk(t)$ is the initial base risk level at time t. B_r is the initial base risk level, D_a is the days in arrear, L_a is the total age of the loan in months, and α, β are the scaling exponents (scaling factors) that represents the sensitivity of risk arrears and loan respectively.

3.2 Model evaluation metrics

This section evaluates the performance and accuracy of the models. This helps to measure the ability of the predictions, understand the probability distribution, classify data correctly, and improve the model's reliability and robustness.

These are achieved with the F1 score, the Receiver Operating Characteristic, and the Area Under the Curve (ROC-AUC). The F1 score is given in equation 3.20 as

$$F1 = \frac{2pr}{p + r} \tag{12}$$

with,

$$p = \frac{TP}{TP + FP} \tag{13}$$

and,

$$r = \frac{TP}{TP + FN} \tag{14}$$

Where the accuracy is given by

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

The ROC-AUC score is given as the integral of the TP Rate (TPR) with respect to the FP Rate (FPR) in Equation (16) as:

$$\int_0^1 TPR\left(FPR^{-1}(x)\right) dx \tag{16}$$

From the above, p is the precision, r is the recall, TP is true positives (correctly predicted defaults), TN is the true negatives (correctly predicted non-defaults), FP is the false positives (non-defaults wrongly predicted defaults), and FN is the false negatives

(defaults wrongly predicted as non-defaults). Also, x represents the varying decision thresholds of the classifiers.

It is worth noting that there is a notable class imbalance in the dataset, with non-defaults making up 85% of the total sample and defaults making up only 15%. Since the high headline score (91%) mostly reflects the model's ability in managing the dominant non-default class rather than its marginal success in detecting the minority class, this distribution may make overall accuracy a little distorted indicator. Additionally, a larger False Positive (FP) Rate and the intrinsic challenge of correctly classifying less frequent positive examples are shown by the lower F1 score for defaults (0.82) compared to non-defaults (0.93).

3.3 Ethical considerations

This study strictly adhered to ethical principles. Permission to access loan records was obtained from the rural bank's management, and informed consent was secured. All data was anonymized to protect client identities, and confidentiality was maintained through encrypted storage and restricted access. The study received ethical approval from the Takoradi Technical University Ethics Review Board, ensuring compliance with institutional and national standards. Participants' rights were respected throughout the research process.

4. Results and Discussion

This section reports on the sources of data and the discussions of the descriptive nature of the data for this study

4.1 Data description

The dataset consists of 25,981 loan accounts from a rural bank, featuring borrower profiles, repayment behavior, regulatory classification, loan characteristics, and default status. The dataset includes both categorical and numerical variables.

From Table 1 across the 25,981 loan records, the mean approved (and disbursed) facility size was approximately GHS 6,528 (SD = 15,199), although values ranged widely from 200 to about GHS 795,325. Contractual tenor averaged 11.44 months (SD = 16.11) with a maximum of 289 months, indicating a heavily right-skewed distribution. The typical loan was young: the median loan age was 4 months, well below the median contractual duration of 6 months. Days-in-arrears was highly skewed (M = 337, SD = 840) but the median was 0, reflecting the predominance of current accounts.

Table 1. Descriptive statistics for continuous variables

Variable	N	M	SD	Min	Max
Loan Amount Approved	25981.0	6528.5	15199.75	100.00	795324.81
Loan Amount Disbursed	25981.0	6528.5	15199.75	100.00	795324.81
Number of Payments Agreed	25981.0	11.44	16.11	0.00	289.0
Loan Principal Balance (Without Interest)	25981.0	4512.0	21526.32	-3000.0	3010616.51
Days in Arrears	25981.0	337.06	840.26	0.0	3248.0
BoG Provision	25981.0	253.29	1194.11	0.0	52559.64
Interest Received	25981.0	11.89	251.9	0.0	18915.11
loan_status_code	25981.0	0.67	1.43	0.0	4.0
loan_duration_months	25981.0	21.71	28.52	-2.0	292.0
loan_age_months	25981.0	16.02	30.39	0.0	1511.0
is_individual	25981.0	0.97	0.18	0.0	1.0

In Table 2 most facilities were held by individuals (96.5%), while a smaller corporate segment accounted for 3.5%. Borrowers were primarily male (68.5%), with females representing 22.5% and 8.9% unspecified. The book concentrated on the "Miscellaneous" economic sector (77.2%), followed by services (11.7%) and commerce/finance (4.9%). Regarding credit quality, 80.3% of exposures were classified as "Current," whereas 14.1% were in "Loss," with the remainder distributed among "OLEM," "Doubtful," and "Sub-standard."

Table 2. Frequency distribution of selected categorical variables

Variable	Category	N	%
Customer Type	Individual	25082	96.5
Customer Type	Corporate	899	3.5
Gender	Male	17805	68.5
Gender	Female	5851	22.5
Gender	Unspecified	2325	8.9
Economic Sector	Miscellaneous	20060	77.2
Economic Sector	Service	3033	11.7
Economic Sector	Commerce & Finance	1267	4.9
Economic Sector	Agriculture Forestry & Fishing	1099	4.2

Variable	Category	N	%
Economic Sector	Manufacturing	325	1.3
Economic Sector	Construction	191	0.7
Economic Sector	Cottage Industries	6	0.0
BOG Loan Classification	Current	20869	80.3
BOG Loan Classification	OLEM	559	2.2
BOG Loan Classification	Sub-Standard	419	1.6
BOG Loan Classification	Doubtful	467	1.8
BOG Loan Classification	Loss	3667	14.1

The target variable is 'is default', a binary label, coded as 1 = Yes, indicating that the loan eventually defaulted, while 0 = No indicates non-defaulted. The analysis revealed that 4,000 loans (approximately 15% of the total sample) were classified as defaults. This suggests that most (85%, 21981) loans remained non-defaulted.

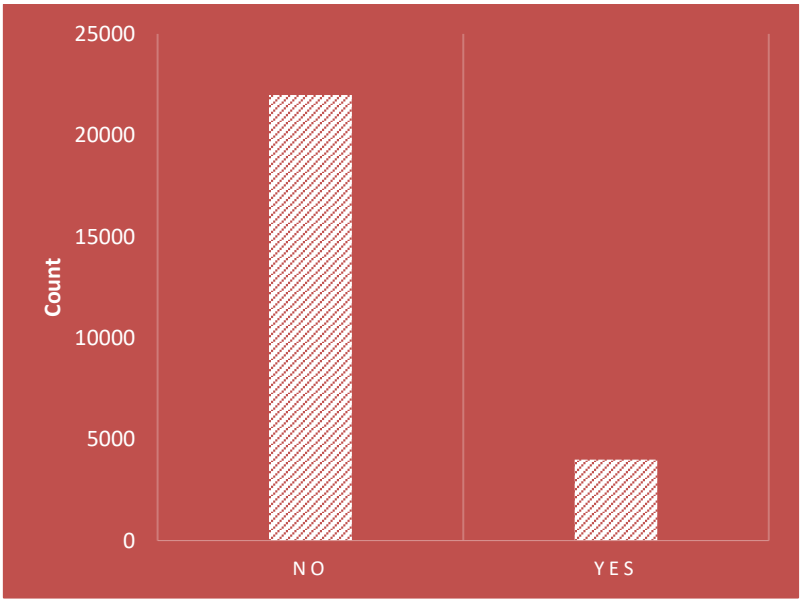

Figure 1 Graph of Default Versus Non-Default

4.2 Markov Chain modeling of Loan status

Loan repayment behavior was modeled as a discrete-time Markov Chain, where each state represents a BOG loan classification: S = {Current, OLEM, Sub-Standard, Doubtful, Loss}.

4.2.1 Distribution of Bank of Ghana (BoG) Loan classification.

Figure 2 shows the percentage of long-term loans in each category, assuming the process continues indefinitely. Current loans (80.32%) make up most of the loan portfolio, indicating good repayment performance. The presence of loans classified as OLEM (2.15%), Sub-standard (1.61%), Doubtful (1.80%), and Losses exceeding ten percent (14.11%) suggests that this Rural Bank's portfolio has some credit concerns.

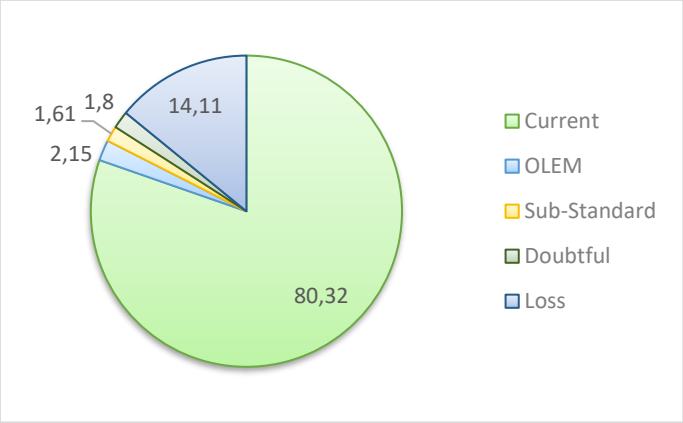

Figure 2: Pie Chart of the Steady State Probabilities (%)

4.2.2 Markov Chain transition Probability Matrix

The transition matrix displaying the probabilities of moving from one state or loan status to another within a single time step is illustrated as a heatmap, as shown in Figure 4.3. In this figure, darker blue represents a higher probability.

Figure 3 shows how the Bank's loans move between risk categories over time. Loans in the current category are highly stable - about 94% of the loans stay current from one month to the next. Only a small fraction slips into early warning or risky categories like OLEM or Loss, a good sign for the bank's overall loan portfolio health.

Interestingly, loans in the OLEM and Sub-Standard categories show a surprisingly high likelihood of reverting to current status, at nearly 74% and 63%, respectively. This may be due to borrowers catching up on payments, stringent monitoring, follow-ups with borrowers, or administrative updates that reclassify loans. However, these loans are not without risk—approximately 16.6% of OLEM loans and 15.8% of Sub-Standard loans still deteriorate into more severe categories such as Doubtful and Loss.

Figure 3 also indicates that the outcomes are more mixed among loans already classified in the Doubtful category. Just over half (56%) recover to current status, which seems optimistic, but a significant portion—more than 11%—ultimately falls into the Loss category, which is concerning.

Finally, loans in the Loss category rarely recover- over 90% stay in the Loss status, confirming that once a loan is written off, it seldom returns.

The diagonal elements in Figure 3 represent the probability of remaining in the same state. For example, the probability of remaining in the "Current (94.47%)" indicates a high chance of staying current, as reflected by the probability of remaining in the "Current" state of 0.9447. This implies that loans currently making their payments on schedule are likely to continue doing so. Furthermore, the chance of remaining Doubtful at 15.42% is comparatively low, with a probability of 0.1542. This suggests that deemed questionable loans are more likely to transition to different states, such as current or loss. In contrast, the probability of remaining in Loss is 90.92%, indicating a strong likelihood (0.9092) of staying in the "Loss" state. This implies that loans declared uncollectible are probably going to remain in this condition. Additionally, the probability of continuing to be in OLEM (9.48%) is relatively low at 0.0948. This suggests an increased likelihood of loans categorized as "OLEM" changing to another state. Finally, the diagonal point for Sub-Standard (11.22%) indicates a reasonable chance of remaining in the "Sub-Standard" state, with a probability of 0.1122. This further suggests that the likelihood of loans categorized as "Sub-Standard" moving to another state is slightly higher.

By comparing the diagonal elements shown in Figure 3, we see a greater degree of stability in the "Current" and "Loss" states, as evidenced by the significantly higher odds of remaining in these states compared to the others. Given the relatively low odds of staying in the "Doubtful," "OLEM," and "Sub-Standard" states, it is likely that loans in these categories will eventually transition to another state.

Further insight into the diagonal entries of TPM, as shown in Figure 6, reveals that the high likelihood of remaining in the "Current" stage suggests that loans currently up to date are repaying their debts effectively. Since these loans are more likely to transition to another state, the comparatively low odds of remaining in the "Doubtful," "OLEM," and "Sub-Standard" states indicate the need for close monitoring. Meanwhile, the high probability of remaining in the "Loss" condition highlights how crucial it is to manage credit risk effectively to minimize the chance of loan default.

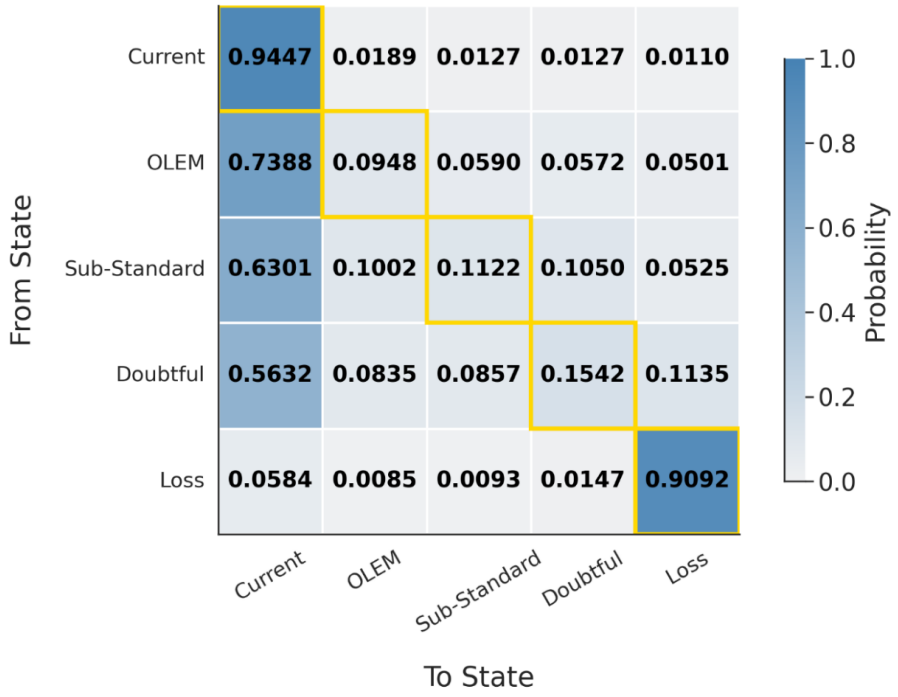

Figure 3 Heatmap of the Markov Chain Transitional Probability Matrix (TPM)

4.2.3 Cluster Analysis

The purpose of the clustering analysing in this research framework is to enable the bank to concentrate its monitoring and intervention strategies on particular segments that have similar repayment tendencies by using the K-Means clustering analysis to group loan statuses according to the similarity of their transition patterns.

Table 3 shows the cluster assignments for the K-Means. The cluster (3-Means cluster) analysis helps the Bank focus on similar-behaving groups of loans, making it easier to design monitoring and intervention strategies.

Table 3 is based on the similarity of the movement patterns of the loan status categories. The loan statuses have been categorized into clusters. “Current” is in group (2) because of its unique behavior from the others. However, "Doubtful," "OLEM," and "Sub-Standard" are comparable and lumped together into one cluster (0), while "Loss" is also distinct and clustered as (1).

Table 3: Cluster Assignment of the K-Means

Loan Status	**Cluster**
Current	2

Loan Status	Cluster
Doubtful	0
Loss	1
OLEM	0
Sub-Standard	0

4.2.4 Customers Risk Scores

Again, Figure 4 displays customers' risk scores according to their present loan status. The customer risk score is the probability that a loan in a given current status will transition to any of the "worse" (risk) states in the next period. Figure 4 shows the likelihood of a customer with "Current, OLEM, Sub-Standard, Doubtful, and Loss" statuses deteriorating as 5.53%, 26.12%, 39.16%, 43.68%, and 94.16%, respectively. This implies that a customer in "Current" status with a risk score of 0.055 has a 5.5% chance of transitioning to a risk status in the next period.

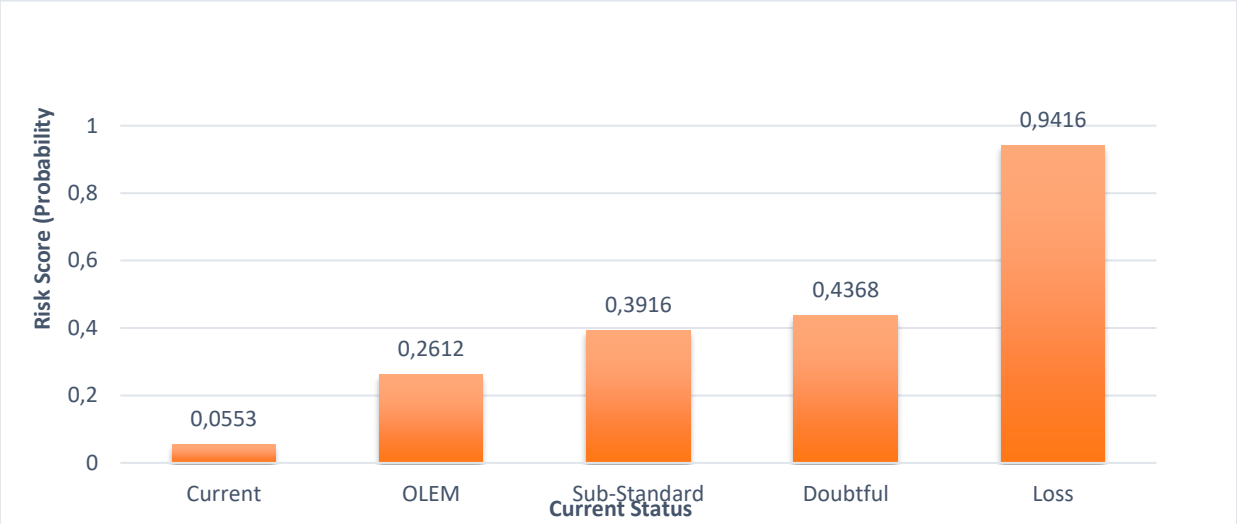

Figure 4: Risk Score by Current Loan Status

Also, Figure 5 shows the expected time to default. Expected time to default is the average number of periods it will take for a customer in a given status to reach the "default" (absorbing) state, assuming the Markov process continues.

Figure 5 shows that, a "Current" loan is expected to take about 69 months before potential default, while an "OLEM" loan has 66 months before potential default, with a "Sub-Standard" loan having about 63 months before defaulting, and a "Doubtful" loan has about 60 months (5 years) before potential default, and

This means that, if a customer's loan moves from "Current" to "OLEM", their expected time to default decreases by about 3 months (from 69 to 66 months).

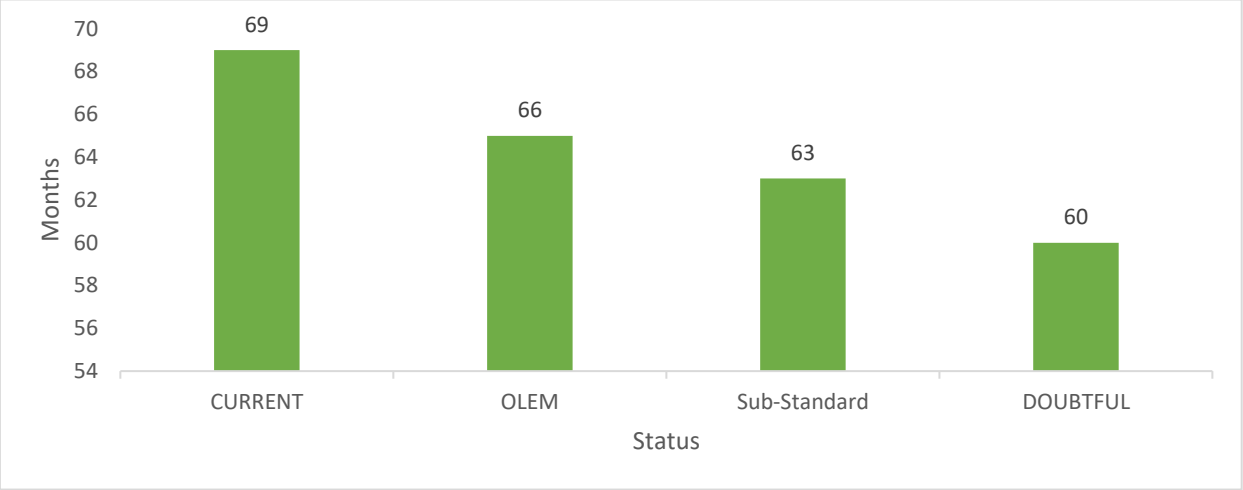

Figure 5: Expected Time of Default (Months)

Combining Figures 4 and 5 reveals the following insights:
1. For loans classified as "Current," the risk of deterioration is 5.53%, with an expected stable period of 5 years and 9 months, necessitating regular monitoring.
2. If the loan transitions to "OLEM," the risk increases to 26.12%, with the expected stable period decreasing to 5.5 years, requiring early intervention.
3. For loans classified as "Doubtful," the risk jumps to 43.68%, with the expected stable period dropping to 5 years, warranting immediate intervention.
These findings highlight the importance of timely risk assessment and intervention to mitigate potential losses.

4.3 Key Feature for Loan Default Prediction.
The objective was achieved by training a Random Forest model and evaluating its performance using various metrics. Default probabilities were simulated under different scenarios, and results were visualized using heatmaps and Receiver Operating Characteristic (ROC) curves.
Table 4 and Figure 6 present the top 10 influential features, accounting for approximately 98% of the model's predictive power. The top features are:
1. Days in arrears (0.300343)
2. IFRS classification (0.154330)
3. BoG loan classification (0.135914)
The least influential feature among the top 10 is the loan principal balance (excluding interest), with a weight of 0.010238.

Table 4: Key Variables Influencing Loan Default Prediction

Feature	**Importance**
Days in Arrears	0.300344
IFRS Classification	0.154330
BOG Loan Classification	0.135914
Interest Charge Type	0.108712
Loan Duration (Months)	0.078287
Loan Age (Months)	0.066854
Frequency of Payments	0.045482
Number of Payments Agreed	0.045155
BoG Provision	0.030630
Loan Principal Balance (without interest)	0.010238

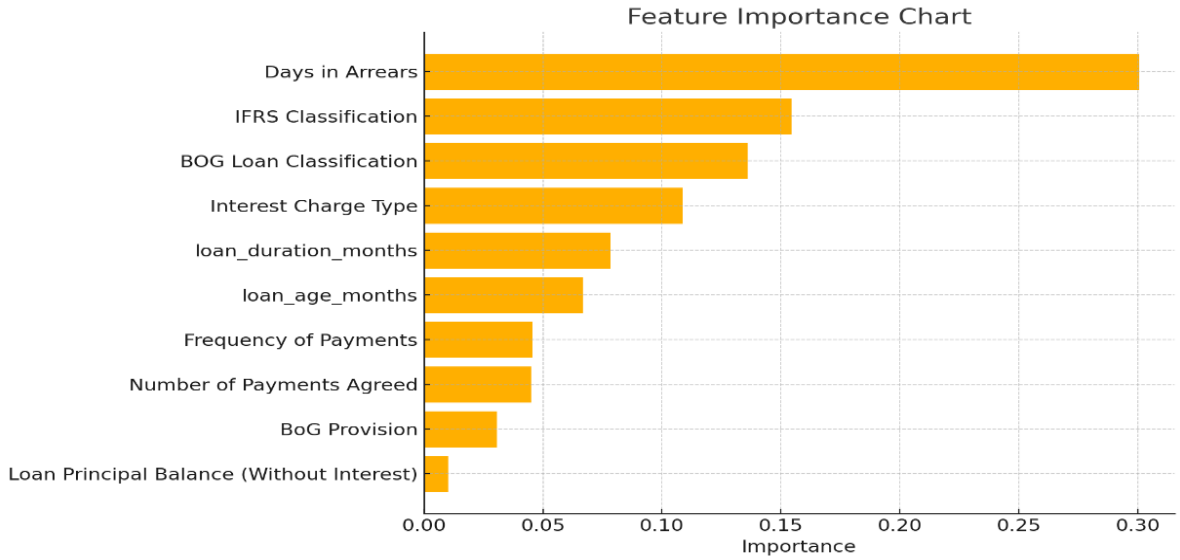

Figure 6: Graph of the Covariate Importance.

4.4 Model Performance Metrics Random Forest

The classifier demonstrated a commendable overall accuracy of 91%, indicating that it correctly classified 91% of the total 5,197 instances in the test dataset (See Table 5 and Figure 7). However, a closer examination of the class-specific metrics reveals nuanced performance across the two categories.

For the negative class (False, n = 4300), the model achieved a precision of 0.94, a recall of 0.93, and an F1-score of 0.93 as indicated on Table 5. These high scores indicate that the model was highly effective in identifying and classifying instances of the negative class. Precision, in this context, refers to the proportion of predicted negatives that were actually correct, while recall represents the proportion of actual negatives that were

correctly predicted. The F1-score, a harmonic mean of precision and recall, confirms that the model maintained a balanced performance for this majority class.

In contrast, the performance metrics for the positive class (True, n=897n=897) were slightly lower. The model recorded a precision of 0.81, a recall of 0.84, and an F1-score of 0.82. Although these values are still indicative of strong performance, they suggest that the model is marginally less effective at identifying the minority class. The lower precision implies a higher false positive rate for the positive class, while the recall rate reveals that the model correctly identified 84% of all true positives.

To address class imbalance and provide a more comprehensive assessment, macro and weighted averages were also calculated. The macro-average values, which treat both classes equally regardless of their frequency, yielded precision, recall, and F1-scores of 0.87, 0.88, and 0.88, respectively. These scores indicate a reasonably balanced performance across the two classes. Meanwhile, the weighted average, which adjusts for the number of instances in each class, reported precision, recall, and F1-scores of 0.91 each—consistent with the overall accuracy and further affirming the model's strength in handling the dominant class.

Table 5: Random Forest Performance Metrics

	Precision	**Recall**	**F1-Score**	**Support**
False	0.94	0.93	0.93	4300
TRUE	0.81	0.84	0.82	897
ROC AUC Score			0.85	5197
Accuracy			0.91	5197
Macro Avg	0.87	0.88	0.88	5197
Weighted Avg	0.91	0.91	0.91	5197

The random forest demonstrates significant additional predictive value, while the ROC AUC offers a more thorough perspective on classification skill at all decision thresholds.

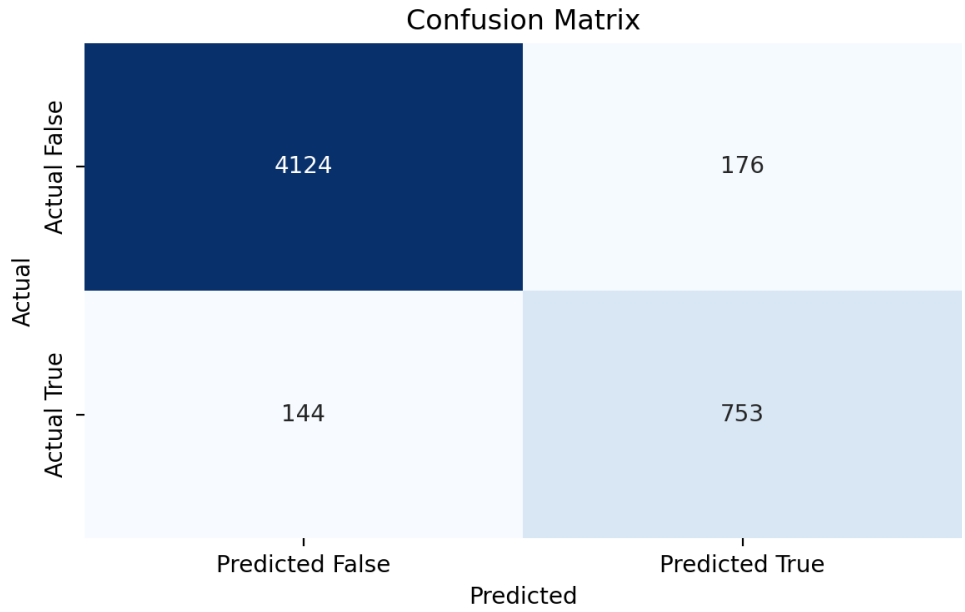

Figure 7: Confusion Matrix of the Random Forest Prediction

Additionally, simulations for different scenarios were conducted using the developed model. Figure 8 displays the scenario simulation heatmap illustrating how the probability of default varies as "Days in Arrears" and "Loan Age (months)" change for a typical customer at the Bank. The darker red in Figure 8 indicates a higher probability of default. As the "Days in Arrears" increases, there is also a sharp rise in the risk of default, especially for loans with a longer duration (age) in months. For instance, the risk increases after 2 months in arrears, but particularly after 3 months in arrears.

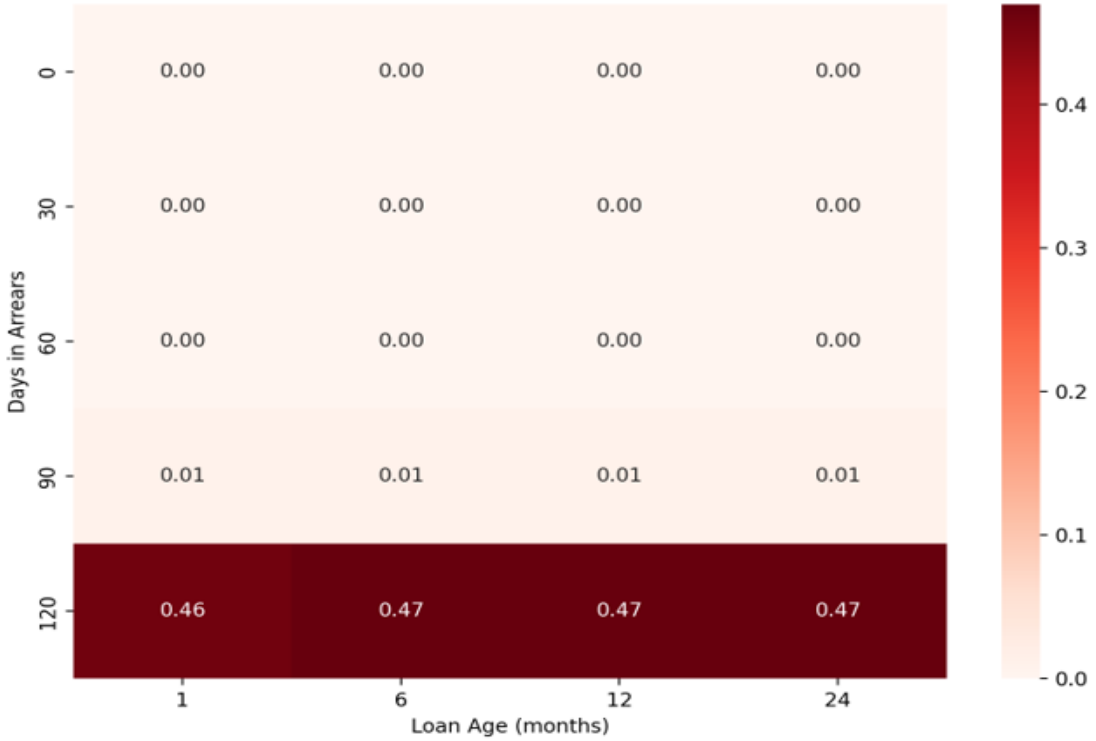

Figure 8: Simulated Default Probability by Days in Arrears and Loan Age (Months)

4.5 Improved Prediction Using Hyperparameter Tuning

From Table 6, after hyperparameter tuning, the random forest model exhibited robust and well-balanced predictive performance. Overall accuracy reached 0.93, indicating that 93% of cases were accurately classified across the ten cross-validation folds. The receiver operating characteristic area under the curve (ROC-AUC) was 0.90, reflecting excellent global discrimination between defaulting and non-defaulting loans (Hosmer, Lemeshow, & Sturdivant, 2013).

Class-specific results show a modest trade-off between majority and minority groups. For non-defaults, precision and recall were 0.95 and 0.94, respectively, resulting in an F1-score of 0.94. For defaults—arguably the more critical category—precision improved to 0.85 and recall to 0.88, yielding an F1-score of 0.86. These values indicate that the tuned model correctly identifies 88% of true defaults while limiting false-positive alerts to about 15%.

Macro-averaged metrics (precision = 0.90; recall = 0.91; F1 = 0.90) further confirm balanced performance when each class is weighted equally, whereas weighted averages (≈ 0.93) mirror headline accuracy by accounting for class prevalence (defaults ≈ 17 %).

Table 6: Optimized Model Performance Metrics

	Precision	**Recall**	**F1-Score**	**Support**
False	0.950	0.940	0.945	4300
TRUE	0.850	0.880	0.865	897
ROC AUC Score			0.90	5197
Accuracy			0.93	5197
Macro Avg	0.90	0.91	0.90	5197
Weighted Avg	0.93	0.93	0.93	5197

The hyper-parameter-optimized random forest model yields a sharply rising default probability curve as payment delinquency extends (see Table 7). Accounts that are current (0–30 days past due) exhibit virtually no risk of default; the estimated PD is 0%. Risk rises significantly once arrears reach 60 days, with an 11.6% chance of default. The slope steepens between 60 and 90 days, more than doubling the PD to 28.5%. At 120 days past due, the probability of default escalates to 85.5%, indicating that loans which are four months in arrears are very likely to become charge-offs.

Table 7: Daily Loan Risk Assessment

Days in Arrears	**Default Probability (%)**
0	0.00
30	0.00
60	11.60
90	28.52
120	85.49

5. Summary, Conclusion and Recommendations

This section considers the summary, the conclusions and the recommendation of the study.

5.1 Summary

The primary goal of this study was to create an integrated credit-risk framework for a rural bank by combining Markov chain clustering analysis with a machine-learning classifier. The developed model from this study results indicated that (a) most loans stayed in the Current state or, once written off, remained in Loss; (b) intermediate states showed significant two-way mobility; and (c) a tuned random forest model achieved high

predictive accuracy (overall = 0.93, ROC AUC = 0.90) while identifying 88% of actual defaults. These findings support previous research suggesting that early-stage delinquency is highly useful for predicting defaults (e.g., Brown & Moles, 2021) and expand the literature by confirming this pattern in the context of a rural bank.

The Markov results further showed that the expected time to default varies significantly by initial grade, decreasing from about 66 months for OLEM accounts to 63 months for Sub-Standard accounts. This gradient highlights the importance of detailed provisioning under IFRS 9, as lifetime expected credit losses differ greatly across BoG buckets. Additionally, scenario simulations based on the tuned random forest revealed a non-linear increase in default probability once arrears exceeded 60 days (PD = 11.6%) and nearly certain loss at 120 days (PD = 85.5%). These time-based insights stress that the best intervention window is before the 90-day mark, when default probability rises sharply.

Additionally, we created a Markov-chain map that illustrates the progression of loans from "Current" to "Loss". The developed model gives a realistic and practical guide for action. The model additionally provided a coherent, data-driven solution for rural bank credit management.

The deductions made from the results of the developed model are; most loans are positioned at the secure extremes of the spectrum. These loans either remain healthy or, in cases of default, they become difficult to recover. The perilous zone is found in the middle. This implies that, around the two-month overdue threshold, the likelihood of complete default significantly increases. The model provided a framework that enhances the bank's risk-management capability beyond conventional rule-of-thumb methods**.** This enables the manager to identify high-risk customers that had to be contacted over time without the manager relying on gut feelings. The model further unraveled that early outreach to customers, a payment plan offer, and a reminder through any form of communication sent 30 to 60 days past due are very crucial in the bank's strategy to mitigate the risk of default by the customers.

In addition, there are three indicators of concern in this study, which were the primary causes of default. These primary causes are, the days past due, the official IFRS stage, and the Bank of Ghana's designated loan-quality classification.

Finally, the model showed an accuracy level of 93%, indicating how efficient the model could be applied in the banking sector by capturing the long-term transition dynamics and the borrower-level risk scores, which are suitable for daily decision-making.

5.2 Conclusions

In this study, a hybrid Markov Chain and Random Forest model for loan default in rural banks in Ghana have been developed. Also, the feature importance of the model has been determined to be ten major variables including, days in arrears, IFRS classification, BoG

classification, interest charge type, etc. Finally, the hyperparameter tuning of the model has been established.

5.3 Practical Implications

From the outcomes of the study, the following are observed as the practical implications associated to the study. On provisioning; the transition-based expected-loss estimates help improve the accuracy of Stage 2 and Stage 3 classifications, lowering the chances of under- or over-reserving. Also, on the early-warning triggers; it was seen that the tuned classifier achieves 88% recall for defaults, enabling automated alerts when a borrower's predicted PD exceeds a management-defined threshold (e.g., 20%). Moreover, on resource allocation; collection efforts can be tiered for low-touch reminders for 0–30 days overdue, personalized outreach at 31–59 days, and full workout escalation at 60 days or more. Finally, on pricing and product design; the study unraveled that, risk-based interest-rate add-ons can be calibrated using point-in-time PDs generated by the random forest, promoting a more sustainable credit portfolio.

5.4 Recommendations for Future Research

Researchers might expand this work by (a) incorporating macroeconomic leading indicators into the feature set to enable macro scenario forecasting, (b) comparing the Markov–random forest hybrid with gradient boosting or deep learning alternatives, and (c) conducting a multi-bank study to evaluate model transferability across various rural institutions. Additionally, exploring explainable AI techniques (e.g., SHAP values) could help bridge the gap between model accuracy and regulatory transparency.

References

[1] Abraham, H. R. (2023). Appraisal Discrimination: Five Lessons for Litigators. *SMU L. Rev.*, *76*, 205.

[2] Agyapong, D. (2023). *Implications of Monitoring and Evaluation Systems for Small and Medium-Sized Enterprises in Selected Metropolis in Ghana Implications of Monitoring and Evaluation Systems for Small and Medium-Sized Enterprises in Selected Metropolis in Ghana* (Doctoral dissertation, University of Cape Coast).

[3] Brown, K., & Moles, P. (2014). Credit risk management. *K. Brown & P. Moles, Credit Risk Management*, *16*.

[4] Ekpu, V., & Paloni, A. (2016). Business lending and bank profitability in the UK. *Studies in Economics and Finance*, *33*(2), 302-319.

[5] Golin, J., & Delhaise, P. (2013). *The bank credit analysis handbook: a guide for analysts, bankers and investors*. John Wiley & Sons.

[6] Oseni, E. (2023). Assessment of the five cs of credit in the lending requirements of the Nigerian commercial banks. *International Journal of Economics and Financial Issues*, *13*(4), 58.
[7] Skiba, P. M., & Tobacman, J. (2008). Payday loans, uncertainty and discounting: Explaining patterns of borrowing, repayment, and default. *Vanderbilt Law and Economics Research Paper*, (08-33).
[8] Vermoesen, V., Deloof, M., & Laveren, E. (2013). Long-term debt maturity and financing constraints of SMEs during the global financial crisis. *Small Business Economics*, *41*(2), 433-448.

Ratio Mathematica Volume 56, 2026

Attractiveness bias in Artificial Intelligence systems

Maria Grazia Olivieri*
Luca Iavarone†

Abstract

Cognitive bias are thought patterns that can lead to incorrect decisions and judgments. Recently, Large Language Models (LLMs) have been shown to be capable of processing and understanding language well. However, because they learn from human data, they may also pick up human bias. While several scholars have analyzed social bias in LLMs, cognitive bias have not been studied as thoroughly. The goal of this research is to analyze how Artificial Intelligence (AI) may be susceptible to a particular cognitive bias, the attractiveness bias, highlighting the consequences of using LLMs. This study adopts an exploratory, simulation-based approach: rather than analyzing real-world court data, it leverages a Large Language Model to generate a synthetic dataset of 500 defendant profiles and corresponding sentences under controlled conditions. This is a bias according to which people considered more attractive are viewed more positively than those considered less attractive. The use of AI in sensitive fields, such as the legal sector, where neutrality is essential, requires regulations that limit the risk of bias. However, the results show that AI technologies are susceptible to cognitive bias, highlighting a lack of clarity and robustness at key decision-making moments.

Keywords: decision analysis, LLM, cognitive bias, attractiveness bias.‡

2010 AMS subject classification: 91B06, 91E10

* Ecampus University, Novedrate, Italy; mariagrazia.olivieri@uniecampus.it
† Ecampus University, Novedrate, Italy; lucaiavarone1992@g.mail.com
[1] Received on March 2, 2026. Accepted on May 31, 2026. Published on June 30, 2026.
DOI: 10.23755/rm.v56i0.1745. ISSN 1592-7415; e-ISSN 2282-8214.
©Olivieri and Iavarone et al. This paper is published under the CC-BY licence agreement.

1. Introduction

Personal choices are increasingly influenced by mental bias, cognitive shortcuts that often result in irrational decisions in interpreting reality [14]. It is believed that using AI to analyze data objectively can play a key role in addressing cognitive bias, rather than relying on a rational person. However, bias present in large language models (LLMs) have emerged as a serious and multifaceted problem, as they can reflect pre-existing human prejudices and social injustices [10]. These models can amplify such perceptual distortions. Bias can arise from various factors, including the sampling of training data, errors in algorithmic performance, and cognitive bias. The paper examines attractiveness bias within the justice system, focusing on its impact on criminal justice decisions. It highlights the general tendency of decision-makers to associate positive qualities with those they consider attractive, while identifying unsuitable qualities in those who do not conform to prevailing aesthetic standards [19]. This is due to our brain's tendency to simplify reality by classifying people into predefined patterns. In the justice system, for the same crime [13], individuals with a more attractive physical appearance have received more favorable treatment than those considered less attractive. Attractiveness bias is a form of the halo effect, a mental process that facilitates complex decisions through visual simplifications [4]. The article explores the psychological processes underlying this bias in the judicial process, particularly the impact on consistency that can arise from implicit bias, which can lead judges or jurors to be unconsciously influenced by irrelevant factors such as the physical appearance of the defendant. LLMs, having been trained on large text corpora, reflect the linguistic patterns found in the underlying data. Since human-written texts often convey implicit bias related to physical appearance, AI algorithms also have the ability to assimilate and reproduce these associations [18]; that is, they replicate the linguistic and conceptual patterns present in the material on which they were trained. Indeed, despite lacking the ability to perceive visual beauty, they can learn, emulate, and reinforce unfair links between beauty and moral, legal, or social evaluations. This implies that, even in the absence of visual input, terms like "nice" or "scruffy" invoke learned bias, thereby influencing the interpretation of legal contracts, ethical assessments, or decision-making. Therefore, if data presents cognitive bias, AI absorbs and replicates them [11]. The result is a model that enhances cognitive bias rather than combating them. Thus, the consistency principle acts as a mental barrier, supporting the narrative that "the system is fair" or "the rules apply to everyone without distinction," even in the face of evidence to the contrary. This human tendency toward bias has crucial implications for artificial intelligence. The principle of coherence is one of the many ways in which the human mind seeks stability, but at the expense of truth. In the context of justice, it can legitimize glaring inequalities. In the context of artificial intelligence, it can blind us to the bias inherent in the systems we use daily [21]. This article presents an exploratory analysis of LLM-generated court sentencing data. The study is simulation-based: both the defendant profiles and the sentences are synthetically produced by a Large Language Model under controlled prompt conditions. Its aim is not to draw causal inferences from real-world judicial data, but rather to investigate whether systematic bias patterns emerge in LLM outputs when attractiveness-related information is introduced. An analysis of 500 people in three judgment contexts (without information on physical appearance, labeled

"Attractive," and labeled "Unattractive") shows significant statistical bias. The findings indicate that any reference to a defendant's physical attractiveness, regardless of whether it is framed positively or negatively, leads to significantly longer sentences compared to situations in which no information about appearance is provided. Although defendants described as "unattractive" receive the highest average sentences, the difference between the "attractive" and "unattractive" conditions does not reach statistical significance. This suggests that the principal source of bias lies in the mere salience of physical appearance rather than in its specific positive or negative connotation. Furthermore, the analyses identify prior criminal record as the most influential predictor of sentencing severity, confirming its central role in judicial decision-making. At the same time, the results reveal the presence of a significant gender bias, as male defendants consistently receive longer sentences than female defendants across all experimental conditions. The paper is structured as follows: Section 2 presents the state of the art of the main literature on the topic; Section 3 defines the methodology used; Section 4 reports the simulation using LLMs; Section 5 contains the results and discussions, as well as mitigation strategies and limitations; Section 6 concludes the paper.

2. Literature review

AI systems can unknowingly absorb bias in data, which can create potential disadvantages such as unfair treatment or unequal distribution of services. Furthermore, AI may fail to correctly interpret non-textual information, such as images or sounds. This can lead to incorrect decisions and unreliable results. Cognitive bias, particularly attractiveness bias, have been the subject of numerous studies in the field of rational decision-making, i.e., the decision-making process. However, the literature on the impact of cognitive bias on LLMs is essentially scarce. [9] conducted a study on the relationship between attractiveness and the halo effect called "What is beautiful is good." Sixty University of Minnesota students, half male and half female, participated in the experiment. Each subject was given three different photographs to examine: one of an attractive individual, one of an individual of average attractiveness, and one of an unattractive individual. Participants were asked to judge the subjects in each photograph by choosing from 27 different personality traits (including altruism, conventionality, self-assertion, stability, emotionality, trustworthiness, extroversion, kindness, and sexual promiscuity). The results showed that the vast majority of participants believed that the most attractive individuals had more socially desirable personality traits than those who were either averagely attractive or unattractive. Participants also believed that attractive people generally led happier lives, had happier marriages, were better parents, and had more successful careers than those who were either unattractive or less attractive. [5] explore in their study the extent to which Artificial Persons (APs) generated by Large Linguistic Models (LLMs) may exhibit cognitive bias similar to those observed in humans. Their findings reveal that APs can replicate these bias, often exaggerating their effects and consequences; they behave as caricatures of human cognitive behavior. [16] in their study, using a set of representative small to medium sized LLMs covering different forms of bias, from physical characteristics to socioeconomic categories,

provide a unified benchmark assessment. The researchers propose five prompting approaches to perform the bias detection task on different aspects of bias. The results indicate that each of the selected LLMs exhibits one or more forms of bias. [10] introduce BiasBuster, a framework designed to uncover, assess, and mitigate cognitive bias in LLMs, particularly in high-stakes decision-making tasks. They develop a dataset containing 13,465 prompts to assess LLMs' decisions based on different cognitive bias and test several bias mitigation strategies. The analysis provides a comprehensive picture of the presence and effects of cognitive bias in commercial and open-source models. [6] in their research asked a series of LLMs to emulate or advise on people's decisions in realistic moral dilemmas. In collective action problems, the LLMs were found to be more altruistic than the participants. In moral dilemmas, the LLMs showed a stronger omission bias than the participants. The results suggest that uncritical reliance on LLMs' moral decisions and advice could amplify human bias and introduce potentially problematic bias. [17] propose a cognitive debiasing approach, self-adaptive cognitive debiasing (SACD), that improves the reliability of LLMs by iteratively refining prompts. The method follows three sequential steps - bias determination, bias analysis, and cognitive debiasing to iteratively mitigate potential cognitive bias in prompts. Experimental results on financial, healthcare, and legal decision-making tasks, using both closed-source and open-source LLMs, demonstrate that the proposed SACD method outperforms both state-of-the-art rapid design methods and existing cognitive debiasing techniques in terms of average accuracy in both single- and multi-bias settings. [15] propose a framework based on behavioral economics theories to assess LLMs' decision-making behaviors. Using a multiple-choice experiment, they initially estimate the degree of risk preference, probability weighting, and loss aversion in a free-flowing setting for three commercial LLMs: ChatGPT-4.0-Turbo, Claude-3-Opus, and Gemini-1.0-pro. The results reveal that LLMs generally exhibit human like patterns, such as risk aversion and loss aversion, with a tendency to overweight small probabilities, but there is significant variation in the degree to which these behaviors are expressed across LLMs.

3. Methodology

The analysis is conducted on a dataset comprising 500 defendants, for each of whom a sentence is determined under three distinct experimental scenarios. To examine the data, several statistical procedures are employed, following the methodological approaches outlined by [2]. Descriptive statistics, including means and frequency counts, are first used to provide an overall summary of the dataset and to identify general patterns across the different conditions. Subsequently, paired t-tests are applied to compare sentencing outcomes for the same individuals across the three attractiveness conditions, such as the comparison between the baseline and the attractive condition. Independent t-tests, following the framework proposed by [20], are instead employed to evaluate differences in average sentence lengths between independent groups, for example between male and female defendants. Throughout the analysis, statistical significance is assessed using a conventional threshold of $p < 0.05$.

Since the study involves both within-subject and between-group comparisons, two forms of the t-test were employed.

For paired comparisons - i.e., comparing sentence lengths assigned to the same defendants across different attractiveness conditions - the paired-samples t-statistic (t_p) is computed as [12]:

$$t_p = \frac{\bar{d}}{(s_d \,/\, \sqrt{n})}$$

where $\bar{d}$ is the mean of the within-subject differences $d_i = x_{i,A} - x_{i,B}$, s_d is the standard deviation of those differences, and n is the number of paired observations. The degrees of freedom are $df = n - 1$.

For independent-group comparisons - e.g., comparing sentence lengths between male and female defendants - the independent-samples t-statistic (t_i) is computed using Welch's variant [20]:

$$t_i = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{({s_1}^2/n_1 + {s_2}^2/n_2)}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the group means, ${s_1}^2$ and ${s_2}^2$ are the group variances, and n_1 and n_2 are the respective sample sizes. This variant does not assume equal variances between groups. Degrees of freedom were estimated using the Welch-Satterthwaite approximation. All results are reported with standard deviations (SD) and 95% confidence intervals (CI) for the relevant mean differences.

This t-statistic is then used to determine the p-value [12]. The p-value represents the probability of observing a result as extreme as, or more extreme than, the one obtained, assuming that no real difference exists between the groups [1]. A small p-value (typically < 0.05) indicates that the observed result is unlikely to have occurred by random chance alone; therefore, the difference is considered statistically significant. In contrast, a large p-value (≥ 0.05) suggests that the observed difference could reasonably be attributed to random variation, and the difference is therefore considered not statistically significant.

4. Case study

This section describes the simulation-based experimental design. All data analyzed in this study - both the defendant profiles and the corresponding sentences - are entirely generated by LLMs under controlled conditions. No real-world judicial data are used. Our goal was to explore whether systematic bias patterns emerge in LLM-generated sentencing when attractiveness information is introduced. We have generated 500 mock personalities using the following prompt against Gemini 2.5 PRO (version July 2025):

Criminal record (yes/no)
Type of offense if there is a criminal record
Personality
Age
Class
Profession
Level of education
Marital status
Hobby
Sex
Number of children

This process created the file “criminalsInput.csv” which contains the 500 personalities. Then we have used the colab script in “Python_Script_for_Sentencing_Analysis.ipynb” to generate and run 3 different prompts for each personality against the LLM “gemini-2.5-flash-lite-preview-06-17”. These 3 prompts are almost the same, but with some key differences in the second part: the first prompt part explains to the LLM to impersonate a judge and gives information about the felony committed by the person. It is important to emphasize that the felony is exactly the same for all the 500 personalities.

In the second part we have added personal information and request to provide suggested sentences in years. The only difference in the last part for the 3 prompts is that the first prompt does not indicate whether the defendant is attractive, the second one instead mentions that the defendant is attractive and the third prompt mentions that the defendant is not attractive. The people's information is conveyed with no gender or plural grammatical adjustments as if it is a preformatted form.

5. Results and discussion

Attractiveness

The average sentence length differed between the three groups (Figure 1). The average sentence length in the "Basic Sentence (No Information)" condition was 2.63 years (SD = 0.78), compared to 2.71 years (SD = 0.72) in the "Attractive" condition and 2.74 years (SD = 0.72) in the "Unattractive" condition.

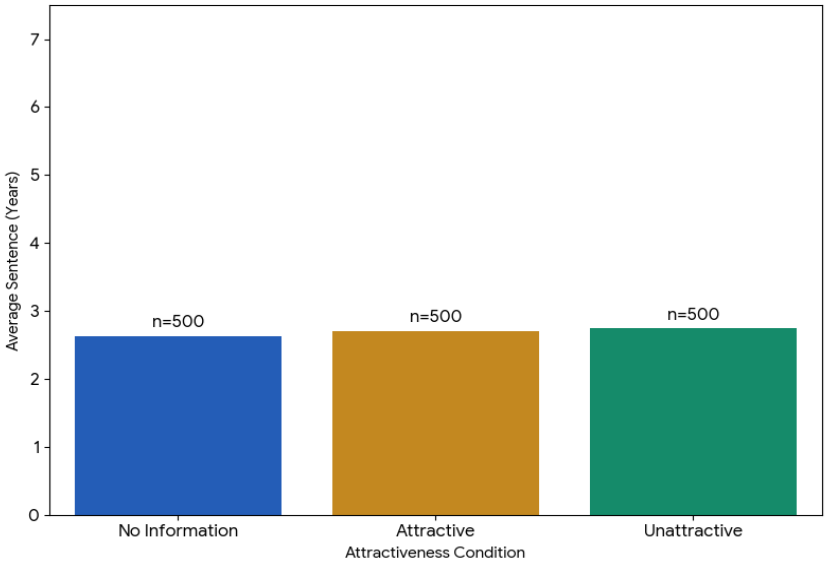

Figure 1. Overall impact of attractiveness information (n=500)

Paired t-tests confirm that these increases are not random. The increase from the Baseline condition to the Attractive condition is statistically significant ($\bar{d}$ = 0.07 years, CI [0.003, 0.141], t(499) = 2.05, p ≈ 0.041), while the increase from the Baseline condition to the Unattractive condition is highly significant ($\bar{d}$ = 0.11 years, CI [0.037, 0.179], t(499) = 3.01, p ≈ 0.003). These results indicate that any mention of appearance leads to harsher judgments. Although labeling a defendant as "unattractive" resulted in a slightly higher average sentence than labeling the defendant as "attractive," this difference is not statistically significant ($\bar{d}$ = 0.04 years, CI [−0.034, 0.106], p ≈ 0.31). Since the confidence interval includes zero and the p-value is much greater than 0.05, we cannot conclude that there is a meaningful difference between the two labels. The primary source of bias therefore appears to arise from making attractiveness salient in the first place.

Criminal record

The analysis found that defendants with criminal records received longer overall sentences than those without. For defendants without a criminal record (n = 473), the average sentence was 2.56 years (SD = 0.66) in the Base condition, 2.64 years (SD = 0.60) in the Attractive condition, and 2.68 years (SD = 0.63) in the Unattractive condition. Conversely, defendants with criminal records (n = 27) received significantly more severe overall sentences, with average sentences of 3.96 years (SD = 1.34) in the Base condition, 3.85 years (SD = 1.38) in the Attractive condition, and 3.89 years (SD = 1.12) in the Unattractive condition. The presence of a criminal record was the most dominant factor influencing sentence length (Figure 2). An independent-samples t-test on the Baseline sentences confirms that this difference is highly significant ($\bar{x}_1 - \bar{x}_2$ = 1.40 years, CI [0.870, 1.939], t(26.7) = 5.39, p < 0.0001).

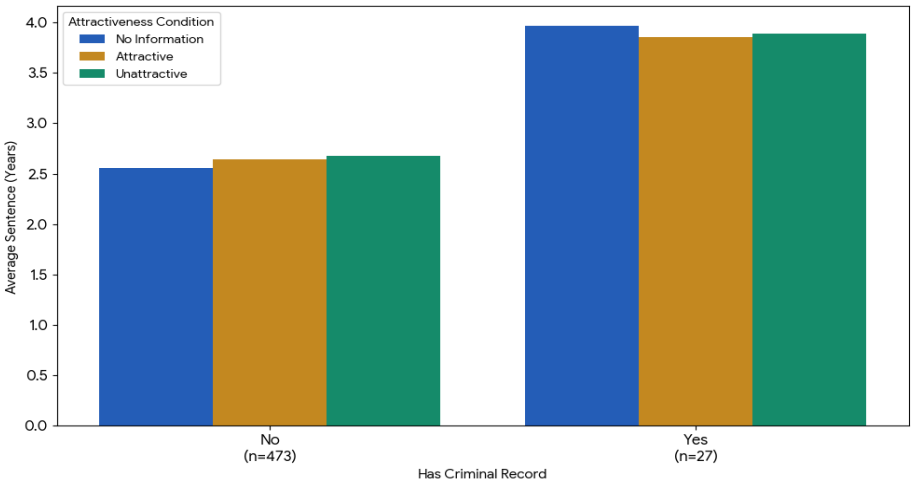

Figure 2. Impact of criminal record and attractiveness on prison sentence

Interestingly, for the 27 individuals with a criminal record, mentioning attractiveness did not increase the sentence, suggesting that a major aggravating factor can overshadow lesser bias.

Gender

For female participants (n = 228), the average sentence was 2.51 years (SD = 0.61) in the Base condition, 2.57 years (SD = 0.59) in the Attractive condition, and 2.59 years (SD = 0.58) in the Unattractive condition. For male participants (n = 272), the corresponding averages were higher across all conditions, with 2.74 years (SD = 0.88) in the Base condition, 2.82 years (SD = 0.80) in the Attractive condition, and 2.87 years (SD = 0.79) in the Unattractive condition. Males consistently received longer sentences than females. A Welch's independent-samples t-test on the Baseline sentences confirms this difference is highly statistically significant ($\bar{x}_1 - \bar{x}_2 = 0.23$ years, CI [0.099, 0.362], t(482.8) = 3.44, p ≈ 0.001).

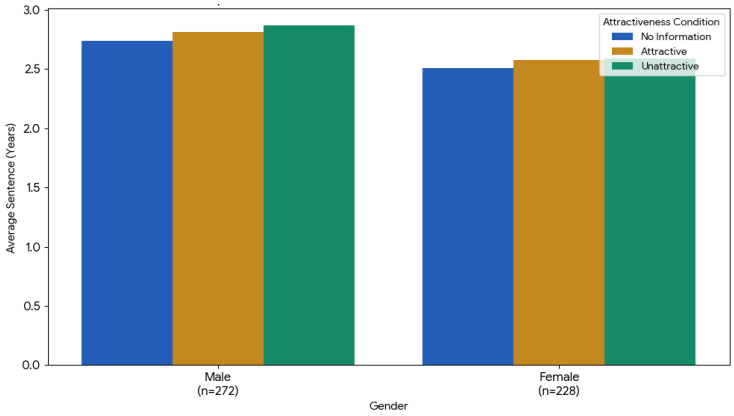

Figure 3. Impact of gender and attractiveness on prison sentence

Social class

With regard to social class, participants in the lower middle class (n = 48) assigned mean sentences of 2.46 years (SD = 0.50) in the Base condition, 2.63 years (SD = 0.49) in the Attractive condition, and 2.60 years (SD = 0.49) in the Unattractive condition. Participants in the upper middle class (n = 77) produced mean sentences of 2.56 years (SD = 0.57) in the Base condition, 2.73 years (SD = 0.66) in the Attractive condition, and 2.75 years (SD = 0.57) in the Unattractive condition. Those in the middle class (n = 228) assigned mean sentences of 2.57 years (SD = 0.68) in the Base condition, 2.63 years (SD = 0.56) in the Attractive condition, and 2.67 years (SD = 0.60) in the Unattractive condition. Participants in the working class (n = 82) showed higher overall sentencing, with mean values of 2.80 years (SD = 0.99) in the Base condition, 2.90 years (SD = 0.99) in the Attractive condition, and 2.90 years (SD = 0.92) in the Unattractive condition. Finally, participants in the upper class (n = 61) assigned mean sentences of 2.98 years (SD = 0.88) in the Base condition, 2.90 years (SD = 0.75) in the Attractive condition, and 3.02 years (SD = 0.87) in the Unattractive condition. Social class presented a more complex interaction. The defendant count for the "Lower Class" was only one, making its result anecdotal (Figure 4).

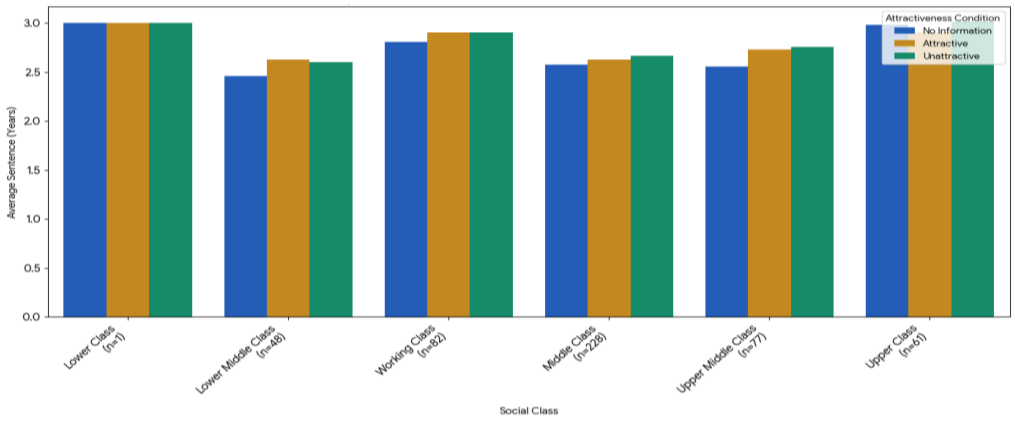

Figure 4. Impact of social class and attractiveness on prison sentence

For most groups, sentences increased when attractiveness was mentioned. The "Upper Class" group showed a unique pattern where being labeled "attractive" was associated with a slight decrease in sentence length. Although this finding is interesting the p-value is 0.55 which makes this correlation not statistically significant.

5.1 Mitigation strategy and limitations

It remains unclear whether the biases observed in large language models (LLMs) primarily arise from pre-existing societal biases reflected in their training data, or whether they are instead largely the result of imperfections in the training process itself. Further research is therefore needed to clarify the extent to which each of these factors contributes to the biases exhibited by LLMs [3]. Also our research is based only on one type of felony

which may involve specific biases that do not generalize to other offenses. A limitation of our research is that we didn't mention to the LLM the gender to impersonate while being a judge: further research could explore if the gender bias depends on the gender of the judge or on if the same/opposite gender is a factor also. Anyway as far as this paper results go, if the goal is to use LLMs to support judges, for example, it could be useful to omit some information like gender or attractiveness. Criminal record might be a legal legit factor which cannot be negated to the LLMs.

If real judges share the same cultural bias as LLMs (which again is yet to be proved) would be hard to completely omit gender and attractiveness. It could make sense to explore a kind of trial where at least in the first iterations the gender/defendant's physical appearance is hidden (no face to face).

Being exploratory and simulation-based, this study has several inherent limitations that should be acknowledged. First, all results are contingent on the specific LLM employed (Gemini 2.5 Flash Lite). Different models, trained on different corpora and with different architectures, may exhibit different bias patterns or magnitudes. The generalizability of the findings to other LLMs - whether commercial or open-source - remains an open question that future replication studies should address. Second, the outcomes may be sensitive to the specific prompt design used to elicit sentences. Variations in prompt wording, structure, or the order in which defendant attributes are presented could yield different results. Although the prompts were kept as neutral and uniform as possible, the absence of a systematic prompt sensitivity analysis means that the robustness of the observed bias across alternative prompt formulations has not been established. Third, this study relies entirely on synthetic LLM generated data and does not include any external or real-world validation. The bias patterns observed in simulated sentencing outputs may not directly mirror those present in actual judicial decisions. Consequently, the findings should be interpreted as evidence of bias tendencies within the LLM's output space, rather than as claims about real-world judicial behaviour. Future work could compare LLM generated sentencing patterns with empirical court data to assess the ecological validity of the simulation based approach adopted here.

6 Conclusion

The central finding of this study is that extraneous information, such as a defendant's physical appearance, can produce statistically significant differences in sentencing outcomes within a fully simulated decision-making environment generated by a large language model (LLM), often resulting in harsher judgments. However, these effects are not uniform and appear to interact with other salient factors within the model's internalised representations. In particular, prior criminal record emerges as the strongest determinant of sentencing severity in the simulated outputs, while a persistent gender-

related effect is also observed, with male defendants receiving consistently higher sentences on average. The principal appearance related effect appears to be driven by the mere salience of attractiveness within the LLM's decision context, as the contrast between "attractive" and "unattractive" labels does not yield a statistically significant difference in sentencing outcomes within the simulated framework. These results reflect patterns in model-generated outputs conditioned on the experimental prompts, and therefore constitute an analysis of bias manifestations in a simulated inference space rather than observations of human decision-making. Within this context, the findings nevertheless underscore the importance of procedural fairness, illustrating how cognitive like biases encoded in model behaviour may influence outputs in ways that diverge from normative principles of equal treatment and justice. From an ethical perspective, this is particularly relevant given the increasing consideration of AI systems for decision-support roles in high stakes domains such as the judiciary, where safeguards are required to mitigate the propagation of biased outputs. The evidence further suggests that, despite advances in model sophistication, LLMs remain vulnerable to bias-related distortions arising from their training data and generative mechanisms, highlighting limitations in their robustness when applied to structured evaluative tasks. Future research should extend this work by incorporating alternative model architectures, systematically varying prompt formulations, and validating simulated sentencing patterns against empirical judicial data in order to strengthen the robustness and external validity of the conclusions.

References

[1] Ackerman, S., Farchi, E., Raz, O., & Toledo, A. (2025). Statistical multi-metric evaluation and visualization of LLM system predictive performance. arXiv preprint arXiv:2501.18243.

[2] Ameli, S., Zhuang, S., Stoica, I., & Mahoney, M. W. (2024). A statistical framework for ranking llm-based chatbots. arXiv preprint arXiv:2412.18407.

[3] Agarwal, D., Fabbri, A. R., Risher, B., Laban, P., Joty, S., & Wu, C. S. (2024, November). Prompt leakage effect and mitigation strategies for multi-turn LLM applications. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track (pp. 1255-1275).

[4] Batres, C., & Shiramizu, V. (2023). Examining the "attractiveness halo effect" across cultures. Current Psychology, 42(29), 25515-25519.

[5] Campbell, H., Goldman, S., & Markey, P. M. (2025). Artificial intelligence and human decision making: Exploring similarities in cognitive bias. Computers in Human Behavior: Artificial Humans, 4, 100138.

[6] Cheung, V., Maier, M., & Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. Proceedings of the National Academy of Sciences, 122(25), e2412015122.

[7] Cordeiro, G. M., Ferrari, S. L., & Cysneiros, A. H. M. (1998). A formula to improve score test statistics. Journal of Statistical Computation and Simulation, 62(1-2), 123-136.

[8] Dash, B. P., & Obaidullah, K. A. (1970). Determination of signal and noise statistics using correlation theory. Geophysics, 35(1), 24-32.

[9] Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. Journal of personality and social psychology, 24(3), 285.

[10] Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in decision-making with LLMs. arXiv preprint arXiv:2403.00811.

[11] Filandrianos, G., Dimitriou, A., Lymperaiou, M., Thomas, K., & Stamou, G. (2025). Bias Beware: The Impact of Cognitive Biases on LLM-Driven Product Recommendations. arXiv preprint arXiv:2502.01349.

[12] Gravetter, F. J., & Wallnau, L. B. (2014). Introduction to the t statistic. Essentials of statistics for the behavioral sciences, 8(252).

[13] Gulati, A., D'Incà, M., Sebe, N., Lepri, B., & Oliver, N. (2025). Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models. arXiv preprint arXiv:2504.16104.

[14] He, J., & Liu, J. (2025). Investigating the Impact of LLM Personality on Cognitive Bias Manifestation in Automated Decision-Making Tasks. arXiv preprint arXiv:2502.14219.

[15] Jia, J. J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-making behavior evaluation framework for llms under uncertain context. Advances in Neural Information Processing Systems, 37, 113360-113382.

[16] Kumar, C. V., Urlana, A., Kanumolu, G., Garlapati, B. M., & Mishra, P. (2025). No LLM is Free From Bias: A Comprehensive Study of Bias Evaluation in Large Language models. arXiv preprint arXiv:2503.11985.

[17] Lyu, Y., Ren, S., Feng, Y., Wang, Z., Chen, Z., Ren, Z., & de Rijke, M. (2025). Cognitive debiasing large language models for decision-making. arXiv preprint arXiv:2504.04141.

[18] Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS digital health, 2(6), e0000278.

[19] Nienhaus, S. (2025). The Impact of Physical Attractiveness and Type of Crime in Judicial Decision Making (Master's thesis, University of Twente).

[20] Ross, A., & Willson, V. L. (2017). Independent samples T-test. In Basic and advanced statistical tests: Writing results sections and creating tables and figures (pp. 13-16). Rotterdam: SensePublishers.

[21] Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large language models are inconsistent and biased evaluators. arXiv preprint arXiv:2405.01724.

[22] Williams, L. J., & Abdi, H. (2010). Fisher's least significant difference (LSD) test. Encyclopedia of research design, 218(4), 840-853.

Ratio Mathematica Volume 56, 2026

Generalised M-Transformation Weibull Distribution: Statistical Properties and Reliability Applications

Francis Kwame Arpoh*
Gabriel Asare Okyere†
Isaac Akpor Adjei‡

Abstract

Many real-world datasets exhibit complex features that classical lifetime models cannot adequately capture. This limitation highlights the need for new distributions with greater flexibility in skewness and kurtosis, as well as improved fit to observed data. In this study, we introduce the Generalised M-Transformation Weibull (GMTW) distribution to meet these needs. We derive its fundamental statistical properties, including the quantile function and reliability measures. The GMTW distribution can represent diverse hazard rate shapes including bathtub, unimodal, and monotone forms—making it highly suitable for reliability studies. Parameters are estimated using the maximum likelihood method, and the asymptotic properties of the estimators are evaluated through extensive Monte Carlo simulations. The model's practical utility is shown using three datasets: industrial reliability, glass fiber tensile strength, and bladder cancer remission times. Likelihood Ratio Tests and goodness-of-fit criteria confirm that the GMTW distribution provides competitive performance relative to its related forms and other well-known lifetime distributions.

*Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; francisarpoh@gmail.com. Corresponding author.

†Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; gaokyere.cos@knust.edu.gh

‡Department of Statistics and Actuarial Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana; iaadjei.sci@knust.edu.gh

Keywords: Weibull distribution; M-transformation; Lifetime distributions; Hazard rate; Reliability analysis[1]
2020 AMS subject classifications: 62N05; 62E15; 62F10

[1]Received on January 1, 2026. Accepted on May 15, 2026. Published on June 30, 2026.
DOI: 10.23755/rm.v56i0.1759. ISSN: 1592-7415. eISSN: 2282-8214.
©Arpoh et al. This paper is published under the CC-BY licence agreement.

1 Introduction

Reliability analysis is a key component in many fields including engineering, medicine, and social sciences. The accuracy of any statistical inference greatly depends on the suitability of the probability model that has been chosen to describe the underlying data. Over the years many models have been proposed to describe lifetime data. The Weibull distribution, introduced by Weibull [1951], has remained the most popular choice among the classical models. This is due to its simplicity and flexibility in modelling data with monotone failure rate. However, many real world data exhibit non-monotone failure rates such as bathtub or modified shapes. A typical example is the "bathtub" shape of failure patterns of electronic components and human mortality. These data mostly show high infant mortality (early life), a constant hazard during midlife, and an increasing hazard due to wear out [Aarset, 1987]. This is when the standard Weibull distribution fails. It is unable to capture these complex hazard behaviours due to its restricted hazard rate function that is either increasing, decreasing, or constant.

Several researchers have proposed modifications and generalizations of the Weibull distribution in order to overcome this limitation. The Transmuted Modified Weibull [Khan and King, 2013], the Exponentiated Weibull [Mudholkar and Srivastava, 1993], and the Flexible Weibull [Bebbington et al., 2007] are notable examples. More recently, a new aspect to lifetime distributions was introduced by Kumar et al. [2017]. They proposed a rational transformation known as the M-transformation that offered a significant improvement in modelling reliability data by creating a geometric mixture of survival functions.

In spite of the success of the M-transformation, there is still the need for even better flexibility to vary the sharpness of the bathtub curvature and control the heaviness of the tails of the distribution. In this research, we propose the Generalised M-Transformation Weibull (GMTW) distribution which incorporates an additional shape parameter α into the M-transformation framework. By providing a more versatile structure, our proposed GMTW model is able to adapt to a wider range of empirical data.

The structural foundation of the GMTW distribution is defined by its cumulative distribution function (CDF), which incorporates the power-extended transformation into the baseline Weibull CDF $F(x; k, \lambda)$, as follows:

Definition 1.1 (GMTW Distribution)**.** *The cumulative distribution function (CDF) of the Generalised M-Transformation Weibull (GMTW) distribution is given by:*

$$G(x; k, \lambda, \alpha) = \frac{F(x)^{\alpha} + F(x)}{1 + F(x)^{\alpha}}, \quad x \geq 0, \quad k > 0,\ \lambda > 0,\ \alpha > 0. \tag{1}$$

The corresponding probability density function (PDF) is obtained by differentiation:

Proposition 1.1 (GMTW Density Function)**.** *The probability density function (PDF) corresponding to Eq.* (1) *is expressed as:*

$$g(x;k,\lambda,\alpha) = f(x;k,\lambda)\cdot\frac{1+\alpha F(x)^{\alpha-1}+(1-\alpha)F(x)^{\alpha}}{\left[1+F(x)^{\alpha}\right]^{2}},\quad x\geq 0, \quad (2)$$

where $f(x;k,\lambda)$ is the baseline Weibull PDF. These two functions form the mathematical core of the model. The additional shape parameter α modulates the baseline density and hazard to accommodate a wide range of regimes, including the critical bathtub, unimodal, and monotone shapes that are central to this study.

The remainder of this study is organized as follows: In Section 2, we define the CDF, PDF, and hazard function of the GMTW distribution. Section 3 explores the mathematical foundations and specific properties of the model, including the quantile function and reliability measures. It follows with a detailed graphical analysis to illustrate the flexibility of the distribution under varying parameter configurations. Section 4 discusses the parameter estimation using the maximum likelihood method and Section 5 presents a Monte Carlo simulation study to evaluate the asymptotic performance of the estimators. In Section 6, the practical utility of the GMTW model is demonstrated through applications to three real-world datasets from industrial, material, and biomedical fields. Finally, Section 7 provides the conclusion and a summary of the key findings of this study.

2 The Generalised M-Transformation Weibull (GMTW) Distribution

In this section, we formally introduce the Generalised M-Transformation Weibull distribution. The model is constructed by applying a generalized rational transformation to a baseline Weibull distribution. This transformation preserves the desirable properties of the Weibull while introducing an additional shape parameter that enables a much wider range of density and hazard shapes, including bathtub, upside-down bathtub, unimodal, and monotone forms, making it particularly suitable for modeling complex lifetime data.

2.1 Baseline Weibull Distribution

Let X be a non-negative continuous random variable following the two-parameter Weibull distribution with shape parameter $k > 0$ and scale parameter

$\lambda > 0$ [Weibull, 1951]. The cumulative distribution function (CDF) and probability density function (PDF) are defined as

$$F(x; k, \lambda) = 1 - \exp\left[-\left(\frac{x}{\lambda}\right)^k\right], \quad x \geq 0, \tag{3}$$

$$f(x; k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^k\right], \quad x \geq 0. \tag{4}$$

The corresponding survival function is

$$S_W(x; k, \lambda) = \exp\left[-\left(\frac{x}{\lambda}\right)^k\right], \quad x \geq 0. \tag{5}$$

2.2 Transformation Function

We first define the transformation function $T : [0, 1] \to [0, 1]$ that generates the proposed family:

Definition 2.1 (M-Transformation Function)**.** *The generalised M-transformation function is defined as:*

$$T(u; \alpha) = \frac{u^\alpha + u}{1 + u^\alpha}, \quad u \in [0, 1], \quad \alpha > 0. \tag{6}$$

This function generalizes the classical M-transformation introduced by Kumar et al. [2017], which corresponds to the special case $\alpha = 1$:

$$T(u; 1) = \frac{2u}{1 + u}. \tag{7}$$

The parameter α acts as a flexibility multiplier, allowing control over tail behavior, skewness, and hazard curvature.

2.3 GMTW Cumulative Distribution Function

The CDF of the GMTW distribution is obtained by composing the transformation T with the baseline Weibull CDF:

$$G(x; k, \lambda, \alpha) = T\big(F(x; k, \lambda); \alpha\big) = \frac{F(x)^\alpha + F(x)}{1 + F(x)^\alpha}, \quad x \geq 0, \quad k, \lambda, \alpha > 0. \tag{8}$$

The support of the GMTW distribution is $[0, \infty)$, and the parameter space is $(k, \lambda, \alpha) \in \mathbb{R}^3_+$.

Substituting the explicit Weibull CDF (3) gives:

$$G(x; k, \lambda, \alpha) = \frac{\big[1 - \exp\big(-(x/\lambda)^k\big)\big]^\alpha + 1 - \exp\big(-(x/\lambda)^k\big)}{1 + \big[1 - \exp\big(-(x/\lambda)^k\big)\big]^\alpha}. \tag{9}$$

2.4 Mathematical Validity of the GMTW Distribution

Proposition 2.1 (Validity of the GMTW Cumulative Distribution Function)**.** *The function*

$$G(x;k,\lambda,\alpha)=\frac{F(x)^{\alpha}+F(x)}{1+F(x)^{\alpha}},\quad x\geq 0,\quad k>0,\ \lambda>0,\ \alpha>0, \tag{10}$$

where $F(x)=1-\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]$ *is the baseline Weibull CDF, is a valid cumulative distribution function for a continuous random variable on* $[0,\infty)$.

Proof. We verify the four essential properties required for any continuous probability distribution on $[0,\infty)$:

1. **Boundary Conditions:** A valid CDF must satisfy:

$$\lim_{x\to 0^{+}}G(x)=0\quad\text{and}\quad\lim_{x\to\infty}G(x)=1.$$

   As $x\to 0^{+}$, $F(x)\to 0^{+}$. Substituting:

$$G(x)\to\frac{0^{\alpha}+0}{1+0^{\alpha}}=0.$$

   As $x\to\infty$, $F(x)\to 1^{-}$. Substituting:

$$G(x)\to\frac{1^{\alpha}+1}{1+1^{\alpha}}=1.$$

   Thus, the boundary conditions are satisfied.

2. **Non-negativity:** For all $x\geq 0$, $F(x)\in[0,1]$. The numerator $F(x)^{\alpha}+F(x)\geq 0$ and denominator $1+F(x)^{\alpha}>0$, so $G(x)\geq 0$.

3. **Monotonicity (Non-decreasing):** Since the baseline CDF $F(x)$ is non-decreasing, it suffices to show that the transformation $T(u;\alpha)$ is non-decreasing on the interval $[0,1]$. The derivative of the transformation is given by:

$$T'(u;\alpha)=\frac{1+\alpha u^{\alpha-1}+(1-\alpha)u^{\alpha}}{(1+u^{\alpha})^{2}}. \tag{11}$$

   The denominator is strictly positive for all $u\in[0,1]$. To establish that $T'(u;\alpha)>0$ for $u\in(0,1)$, we examine the numerator, $N(u)=1+\alpha u^{\alpha-1}+(1-\alpha)u^{\alpha}$, under two cases for the transformation parameter $\alpha>0$:

- **Case 1:** $0 < \alpha \leq 1$**.** In this scenario, $(1-\alpha) \geq 0$. Since all individual terms (1, $\alpha u^{\alpha-1}$, and $(1-\alpha)u^{\alpha}$) are non-negative for $u > 0$, it follows that $N(u) > 0$.
- **Case 2:** $\alpha > 1$**.** We can rearrange the numerator as $N(u) = 1 + \alpha u^{\alpha-1} - (\alpha-1)u^{\alpha}$. Factoring out terms, we observe:

$$N(u) = 1 + u^{\alpha-1}[\alpha - (\alpha-1)u].$$

Since $0 < u < 1$, the term $(\alpha-1)u < (\alpha-1)$. Therefore, $\alpha - (\alpha-1)u > \alpha - (\alpha-1) = 1$. This ensures that the entire expression is strictly greater than 1.

In both cases, $N(u) > 0$, confirming that $T(u;\alpha)$ is strictly increasing on $(0,1)$. Consequently, the GMTW cumulative distribution function $G(x) = T(F(x);\alpha)$ is non-decreasing for all $x \geq 0$.

4. **Right-continuity and Continuity:** Both $F(x)$ and $T(u)$ are continuous on their domains, so their composition $G(x)$ is continuous on $[0,\infty)$, implying right-continuity everywhere.

These properties confirm that $G(x;k,\lambda,\alpha)$ is a valid CDF for a continuous random variable supported on $[0,\infty)$. □

2.5 GMTW Probability Density Function

The PDF $g(x;k,\lambda,\alpha)$ is obtained by differentiating the CDF:

Proposition 2.2 (GMTW Probability Density Function)**.** *The probability density function of the GMTW distribution is given by:*

$$g(x) = \frac{d}{dx}G(x) = f(x;k,\lambda)\cdot T'\big(F(x;k,\lambda);\alpha\big), \tag{12}$$

where the derivative of the transformation function is

$$T'(u;\alpha) = \frac{1+\alpha u^{\alpha-1}+(1-\alpha)u^{\alpha}}{(1+u^{\alpha})^2}. \tag{13}$$

Since $G(x)$ is a valid cumulative distribution function (Proposition 2.1), its derivative $g(x)$ is necessarily a valid probability density function (non-negative and integrates to 1). The detailed verification, using the change-of-variable technique, is provided in Appendix A.

Substituting yields the explicit PDF:

$$g(x;k,\lambda,\alpha) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right] \times \frac{1+\alpha\big[1-\exp\big(-(x/\lambda)^k\big)\big]^{\alpha-1}+(1-\alpha)\big[1-\exp\big(-(x/\lambda)^k\big)\big]^{\alpha}}{\big\{1+\big[1-\exp\big(-(x/\lambda)^k\big)\big]^{\alpha}\big\}^2}, \quad x \geq 0. \tag{14}$$

2.6 Survival and Hazard Functions

Definition 2.2 (Survival Function)**.** *The survival (reliability) function of the GMTW distribution is*

$$S(x;k,\lambda,\alpha) = 1 - G(x;k,\lambda,\alpha) = \frac{\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]}{1+\big[1-\exp\big(-(x/\lambda)^k\big)\big]^{\alpha}}, \quad x \geq 0. \tag{15}$$

Definition 2.3 (Hazard Function)**.** *The hazard rate function (instantaneous failure rate) of the GMTW distribution is*

$$h(x;k,\lambda,\alpha) = \frac{g(x;k,\lambda,\alpha)}{S(x;k,\lambda,\alpha)} = h_W(x;k,\lambda)\cdot\frac{1+\alpha F(x)^{\alpha-1}+(1-\alpha)F(x)^{\alpha}}{1+F(x)^{\alpha}}, \tag{16}$$

where $h_W(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}$ *is the baseline Weibull hazard function.*

The multiplicative distortion factor

$$D(u;\alpha) = \frac{1+\alpha u^{\alpha-1}+(1-\alpha)u^{\alpha}}{1+u^{\alpha}}, \quad u = F(x) \in [0,1]. \tag{17}$$

is the flexibility multiplier responsible for the model's enhanced ability to generate a wide variety of hazard shapes (monotone, bathtub, upside-down bathtub, etc.) depending on the value of α. It is important to distinguish the transformation derivative $T'(u;\alpha)$ in (13) (which features a squared denominator) from this hazard multiplicative distortion factor $D(u;\alpha)$ (which features a single-power denominator).

Lemma 2.1 (Asymptotic Behaviour of the Distortion Factor)**.** *For a fixed* $\alpha > 0$*, let* $D(u;\alpha)$ *denote the distortion factor defined in Equation* (17)*. The boundary behaviour of* $D(u;\alpha)$ *is characterised as follows:*

1. ***Behaviour at the Origin (*$u \to 0^+$*):***

$$\lim_{u\to 0^+} D(u;\alpha) = \begin{cases} +\infty, & 0 < \alpha < 1, \\ 2, & \alpha = 1, \\ 1, & \alpha > 1. \end{cases} \tag{18}$$

2. ***Behaviour at the Upper Boundary*** *(*$u \to 1^-$*):*

$$\lim_{u \to 1^-} D(u; \alpha) = 1 \quad \textit{for all } \alpha > 0. \tag{19}$$

3. ***Regularity:*** $D(u; \alpha)$ *is strictly positive and continuous on the interval* $[0, 1)$*, and differentiable on* $(0, 1)$ *for all* $\alpha > 0$*.*

Proof. Consider the expression for the distortion factor $D(u; \alpha)$.

(i) As $u \to 0^+$, the term u^α vanishes for all $\alpha > 0$. However, the term $\alpha u^{\alpha-1}$ depends critically on the value of α. If $0 < \alpha < 1$, the exponent $\alpha - 1$ is negative, causing $u^{\alpha-1} \to \infty$ and consequently $D(u; \alpha) \to \infty$. If $\alpha = 1$, then $\alpha u^{\alpha-1} = 1$, yielding $D(0; 1) = (1 + 1)/1 = 2$. For the case where $\alpha > 1$, the exponent is positive, leading to $u^{\alpha-1} \to 0$, which results in $D(0; \alpha) = 1$.

(ii) To examine the behaviour as $u \to 1^-$, we set $u = 1 - \epsilon$ where $\epsilon \to 0^+$. Utilising the Taylor expansions $u^\alpha = 1-\alpha\epsilon+o(\epsilon)$ and $u^{\alpha-1} = 1-(\alpha-1)\epsilon+o(\epsilon)$, we substitute these into the numerator:

$$1+\alpha[1-(\alpha-1)\epsilon]+(1-\alpha)[1-\alpha\epsilon]+o(\epsilon) = 1+\alpha-\alpha(\alpha-1)\epsilon+1-\alpha-\alpha(1-\alpha)\epsilon+o(\epsilon) = 2+o(\epsilon).$$

The single-power denominator simplifies as $(1 + u^\alpha) = (2 - \alpha\epsilon + o(\epsilon))$. Thus, the limit is:

$$\lim_{u \to 1^-} D(u; \alpha) = \frac{2}{2} = 1.$$

(iii) The properties of positivity, continuity, and differentiability follow directly from the algebraic structure of $D(u; \alpha)$ and the well-defined nature of the power function u^α on the open interval $(0, 1)$. □

Remark 2.1. *This asymptotic analysis reveals that for any fixed* $\alpha > 0$*, the hazard function satisfies* $h(x) \sim h_W(x)$ *as* $x \to \infty$*. Furthermore, while* $G(x) \to F(x)$ *pointwise as* $\alpha \to \infty$*, this convergence is non-uniform across the support. For any fixed* x*, the hazard function* $h(x)$ *converges smoothly to the baseline Weibull hazard* $h_W(x)$ *as* $\alpha \to \infty$*, ensuring structural consistency with the nested baseline distribution.*

2.7 Special Cases and Limiting Behaviour

The GMTW distribution family is characterized by its ability to encompass or approach several well-known models depending on the value of the transformation parameter α. When $\alpha = 1$, the model simplifies to the M-Weibull distribution introduced by Kumar et al. [2017], where the cumulative distribution function is given by $G(x; k, \lambda, 1) = 2F(x)/(1 + F(x))$. This specific case represents a balanced rational transformation of the baseline Weibull distribution.

As the transformation parameter increases, the GMTW distribution recovers the baseline model. Specifically, as $\alpha \to \infty$, the transformation function $T(u;\alpha)$ approaches the identity function for all $u \in [0,1)$. Consequently, the GMTW distribution converges to the standard Weibull distribution, such that $\lim_{\alpha\to\infty} G(x;k,\lambda,\alpha) = F(x;k,\lambda)$. This property demonstrates that the GMTW distribution can be viewed as a flexible extension that generalizes the classic Weibull framework.

Finally, it is essential to define the domain of validity to ensure the GMTW remains a proper lifetime model. As $\alpha \to 0^+$, the transformation approaches the limiting form $G(x) \to \frac{1+F(x)}{2}$, which implies $G(0^+) = 1/2$. This limiting case represents a defective distribution with a significant point mass at the origin, a characteristic common in zero-inflated phenomena. To ensure the model remains non-defective and assigns no probability mass to the origin ($G(0) = 0$), we strictly define the parameter space as $\alpha \in (0,\infty)$. Under this constraint, $\lim_{x\to 0^+} G(x;k,\lambda,\alpha) = 0$ for all $\alpha > 0$, confirming the properness of the distribution across its intended functional domain.

2.8 Identifiability

The property of identifiability is fundamental for the validity of maximum likelihood estimation, ensuring that distinct parameter values correspond to distinct probability distributions.

Theorem 2.1 (Identifiability of the GMTW Distribution)**.** *The GMTW distribution family is identifiable; that is, if for all $x \geq 0$,*

$$G(x;k,\lambda,\alpha) = G(x;k^*,\lambda^*,\alpha^*), \tag{20}$$

then the parameter vectors are identical: $(k,\lambda,\alpha) = (k^*,\lambda^*,\alpha^*)$.

Proof. Suppose there exist two parameter vectors $\theta = (k,\lambda,\alpha)$ and $\theta^* = (k^*,\lambda^*,\alpha^*)$ such that the cumulative distribution functions coincide for all $x \geq 0$:

$$T(F(x;k,\lambda);\alpha) = T(F(x;k^*,\lambda^*);\alpha^*), \tag{21}$$

where $T(u;\alpha) = \frac{u^\alpha+u}{1+u^\alpha}$ is the rational transformation defined in Equation (6) and $F(x;k,\lambda)$ is the baseline Weibull CDF. We establish equality in three logical stages.

Step 1: Identification of baseline parameters k and λ Consider the asymptotic behavior of the survival function $\bar{G}(x) = 1 - G(x)$ as $x \to \infty$. As $u \to 1^-$, let $u = 1-\epsilon$. A Taylor expansion of the transformation around $u = 1$ yields:

$$T(1-\epsilon;\alpha) = \frac{(1-\epsilon)^\alpha + (1-\epsilon)}{1+(1-\epsilon)^\alpha} = \frac{1-\alpha\epsilon+1-\epsilon+O(\epsilon^2)}{1+1-\alpha\epsilon+O(\epsilon^2)} = 1-\frac{1}{2}\epsilon+O(\epsilon^2).$$

Thus, $1 - T(u;\alpha) \sim \frac{1}{2}(1-u)$ as $u \to 1^-$. Applying this to the GMTW survival function as $x \to \infty$:

$$1 - G(x;\theta) \sim \frac{1}{2}e^{-(x/\lambda)^k} \quad \text{and} \quad 1 - G(x;\theta^*) \sim \frac{1}{2}e^{-(x/\lambda^*)^{k^*}}.$$

The equality in (21) implies that $\frac{1}{2}e^{-(x/\lambda)^k} = \frac{1}{2}e^{-(x/\lambda^*)^{k^*}}$ for all sufficiently large x. Taking the natural logarithm twice of both sides yields:

$$k(\ln x - \ln \lambda) = k^*(\ln x - \ln \lambda^*).$$

Since this linear identity in $\ln x$ must hold for a continuum of values, the coefficients must be identical, forcing $k = k^*$ and $\lambda = \lambda^*$.

Step 2: Identification of the transformation parameter α With the baseline parameters identified, we have $F(x;k,\lambda) = F(x;k^*,\lambda^*)$. Let $u = F(x) \in (0,1)$. Equation (21) simplifies to:

$$T(u;\alpha) = T(u;\alpha^*) \quad \text{for all } u \in (0,1). \tag{22}$$

To show $\alpha = \alpha^*$, we examine the partial derivative of $T(u;\alpha)$ with respect to α:

$$\frac{\partial T(u;\alpha)}{\partial \alpha} = \frac{u^\alpha (1-u)\ln u}{(1+u^\alpha)^2}.$$

Since $\ln u < 0$ for $u \in (0,1)$ and all other terms are positive, we have $\frac{\partial T}{\partial \alpha} < 0$. This implies that $T(u;\alpha)$ is a strictly decreasing function of α. Since the mapping $\alpha \mapsto T(u;\alpha)$ is continuous and strictly decreasing on $(0,\infty)$ for each fixed $u \in (0,1)$, equality for all u implies $\alpha = \alpha^*$.

Conclusion Since $k = k^*$, $\lambda = \lambda^*$, and $\alpha = \alpha^*$, the parameter vectors are identical. Thus, the GMTW distribution family is identifiable. □

3 Statistical Properties

In this section, we derive and discuss the main statistical properties of the Generalised M-Transformation Weibull (GMTW) distribution. These include the raw moments (with a series expansion for computation), the quantile function and associated robust shape measures, the behaviour of the hazard rate function, order statistics, and limiting behaviours under different parameter regimes.

3.1 Raw Moments and Moment-Generating Function

Definition 3.1 (Raw Moments)**.** *The r-th raw moment of a GMTW random variable X is given by*

$$\mu'_r = E[X^r] = \int_0^\infty x^r \, g(x; k, \lambda, \alpha) \, dx, \quad r > -1. \tag{23}$$

Using the substitution $u = F(x; k, \lambda)$, so $du = f(x)\,dx$ and $x = \lambda[-\ln(1-u)]^{1/k}$, we obtain the integral representation

$$\mu'_r = \lambda^r \int_0^1 [-\ln(1-u)]^{r/k} \cdot \frac{1 + \alpha u^{\alpha-1} + (1-\alpha)u^\alpha}{(1+u^\alpha)^2} \, du. \tag{24}$$

This integral has no closed form in general, but it is numerically tractable using adaptive quadrature. To facilitate numerical evaluation, especially for large r or high precision, we expand the denominator using the generalized binomial series [Gradshteyn and Ryzhik, 2014], valid for $|u^\alpha| < 1$, which holds almost everywhere on $[0, 1)$:

$$(1+u^\alpha)^{-2} = \sum_{j=0}^{\infty} (-1)^j (j+1) u^{j\alpha}. \tag{25}$$

Substituting this expansion into (24) yields a series representation:

$$\mu'_r = \lambda^r \sum_{j=0}^{\infty} (-1)^j (j+1) \int_0^1 [-\ln(1-u)]^{r/k} \left[1 + \alpha u^{\alpha-1} + (1-\alpha)u^\alpha\right] u^{j\alpha} \, du. \tag{26}$$

Each integral in the sum can be evaluated numerically or, in special cases, related to known forms involving the gamma function.

Definition 3.2 (Moment-Generating Function)**.** *The moment-generating function (MGF), $M(t) = E[e^{tX}]$, is given by*

$$M(t) = \int_0^1 \exp\left\{t\lambda[-\ln(1-u)]^{1/k}\right\} \cdot \frac{1 + \alpha u^{\alpha-1} + (1-\alpha)u^\alpha}{(1+u^\alpha)^2} \, du. \tag{27}$$

The existence of the MGF depends on the shape parameter k as follows:

1. For $k > 1$, the MGF exists for all real t.
2. For $k = 1$, the MGF exists only for t below some positive threshold.
3. For $0 < k < 1$, the MGF does not exist for any $t > 0$.

When the MGF fails to exist for positive t, the distribution is instead characterised by its raw moments or its characteristic function.

Once the first four raw moments $(\mu'_1, \mu'_2, \mu'_3, \mu'_4)$ are obtained, the higher central moments and shape descriptors are computed using the standard relationships:

- Variance: $\sigma^2 = \mu'_2 - (\mu'_1)^2$
- Skewness: $\gamma_1 = \frac{\mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3}{\sigma^3}$
- Kurtosis: $\beta_2 = \frac{\mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 3(\mu'_1)^4}{\sigma^4}$

3.2 Quantile Function and Robust Shape Measures

Definition 3.3 (Quantile Function)**.** *The quantile function, $Q(p) = G^{-1}(p)$ for $p \in (0, 1)$, is derived by inverting the cumulative distribution function. Setting $G(Q(p)) = p$ leads to the following relation:*

$$\frac{v^\alpha + v}{1 + v^\alpha} = p, \quad \textit{where} \quad v = F(Q(p)). \tag{28}$$

Equation (28) is a nonlinear equation that does not admit a closed-form solution for v in terms of elementary functions. However, given that $G(v)$ is strictly monotonic for $\alpha > 0$, a unique solution for v can be efficiently obtained using numerical root-finding algorithms such as the Newton-Raphson or bisection methods. Once v is determined numerically, the quantile is obtained by applying the inverse of the baseline Weibull CDF defined in Equation (3):

$$Q(p) = F^{-1}(v) = \lambda \left[-\ln(1 - v)\right]^{1/k}. \tag{29}$$

The quantile function facilitates the derivation of robust shape measures that do not depend on the existence of higher-order moments. These diagnostics are particularly valuable when the GMTW distribution exhibits heavy-tailed behavior or when moments are algebraically complex [Hyndman and Fan, 1996]:

Definition 3.4 (Bowley's Skewness Coefficient)**.** *A robust measure of asymmetry based on quartiles as proposed in [Bowley, 1920]:*

$$S = \frac{Q(3/4) - 2Q(1/2) + Q(1/4)}{Q(3/4) - Q(1/4)}. \tag{30}$$

Definition 3.5 (Moors' Kurtosis Coefficient)**.** *A robust alternative to Pearson's kurtosis that utilizes octiles to describe the concentration of probability in the tails and center [Moors, 1988]:*

$$K = \frac{Q(7/8) - Q(5/8) + Q(3/8) - Q(1/8)}{Q(6/8) - Q(2/8)}. \tag{31}$$

These quantile-based measures provide stable diagnostics for the distribution's shape and are insensitive to the influence of extreme outliers, ensuring reliable characterization across the GMTW parameter space.

3.3 Behaviour of the Hazard Function

The hazard rate function $h(x)$ is a critical measure in reliability analysis, representing the instantaneous failure rate at time x. The GMTW distribution offers significant flexibility in this regard, as established by the following structural decomposition.

Theorem 3.1 (Decomposition of the Hazard Function)**.** *The hazard rate function of the GMTW distribution can be expressed as the product of the baseline Weibull hazard and a rational distortion factor:*

$$h(x;k,\lambda,\alpha) = h_W(x;k,\lambda)\cdot D(F(x;k,\lambda);\alpha), \tag{32}$$

where $h_W(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}$ *is the baseline Weibull hazard, and the distortion factor* $D(u;\alpha)$ *is defined as in* (17)*.*

Remark 3.1 (Hazard Shape Classification and Mechanisms)**.** *The geometric profile of* $h(x)$ *is governed by the competing rates of change between the monotone baseline Weibull hazard and the distortion factor* $D(u;\alpha)$*. As established in Lemma 2.1, the distortion factor exhibits distinct monotonicity properties based on the value of* α*. Specifically, for* $\alpha > 1$*,* $D(u;\alpha)$ *is strictly increasing in* u*, providing an upward adjustment to the baseline hazard. Conversely, when* $\alpha = 1$*, the factor* $D(u;1) = 2/(1+u)^2$ *is strictly decreasing. In the case where* $0 < \alpha < 1$*,* $D(u;\alpha)$ *is strictly decreasing in* u *and diverges as* $u \to 0^+$*, creating a significant initial multiplier that decays rapidly.*

The interaction of these components allows the GMTW distribution to encompass a wide variety of hazard shapes. A monotone increasing hazard typically occurs when both components are non-decreasing ($k \geq 1$ *and* $\alpha \geq 1$*), or when* $k > 1$ *and* $\alpha < 1$ *if the growth of the baseline Weibull hazard eventually dominates the initial decay of the distortion factor. A monotone decreasing hazard is predominantly observed when* $k \leq 1$ *and* $\alpha = 1$*, or when both the baseline and the distortion factor are decreasing (*$k \leq 1$ *and* $0 < \alpha \leq 1$*). Furthermore, a unimodal or upside-down bathtub hazard arises when an increasing distortion factor (*$\alpha > 1$*) initially dominates a decreasing baseline (*$k < 1$*), creating a peak before the eventual decline.*

Of particular interest is the ability of the model to achieve a bathtub or U-shaped profile through two distinct structural pathways. Under Mechanism I, a decreasing baseline ($k < 1$*) is combined with an increasing distortion factor*

($\alpha > 1$). In this scenario, the initial hazard is high due to the Weibull shape, decreases, and is eventually forced upward by the rational transformation. Under Mechanism II, which is frequently observed in Aarset-type data, an increasing baseline ($k > 1$) is combined with a small transformation parameter ($0 < \alpha < 1$). Here, the distortion factor creates an artificial "burn-in" period of high mortality that decays rapidly, while the underlying increasing Weibull baseline eventually asserts dominance to produce the U-shape.

A rigorous analytical characterisation involves the derivative of the log-hazard, $\eta(x) = \ln h(x)$:

$$\eta'(x) = \frac{h'_W(x)}{h_W(x)} + \frac{D'(F(x);\alpha)}{D(F(x);\alpha)} \cdot f(x). \tag{33}$$

The sign of $\eta'(x)$ determines the local monotonicity of the hazard, where bath-tub shapes correspond to a single sign change of $\eta'(x)$ from negative to positive. Given that the conditions for such transitions depend on k and α in a non-separable, non-linear manner, a closed-form parametric partition of all possible shapes is analytically formidable. Consequently, we rely on numerical exploration and the empirical evidence from real-world applications to demonstrate this structural flexibility.

3.4 Order Statistics

Proposition 3.1 (Order Statistics PDF)**.** *Let $X_1, X_2, \ldots, X_n$ be a random sample from the GMTW distribution. The PDF of the i-th order statistic $X_{(i:n)}$ is*

$$g_{(i:n)}(x) = \frac{n!}{(i-1)!(n-i)!}\, g(x;k,\lambda,\alpha)\,[G(x;k,\lambda,\alpha)]^{i-1} \times [1-G(x;k,\lambda,\alpha)]^{n-i}, \quad x \geq 0. \tag{34}$$

This expression is fundamental in reliability engineering for analyzing k-out-of-n systems, where the i-th failure time is of interest [David and Nagaraja, 2003].

3.5 Graphical Analysis of the GMTW Model

In this subsection, we visually explore the flexibility of the Generalised M-Transformation Weibull (GMTW) distribution through two sensitivity analyses. Figure 1 illustrates the influence of the transformation parameter α, while Figure 2 demonstrates how the model responds to changes in the baseline shape parameter k.

The influence of the transformation parameter α is explored in Figure 1, which displays the PDF $g(x)$, survival function $S(x)$, and hazard function $h(x)$ for varying values of α while keeping $k = 0.8$ and $\lambda = 1$ constant. The density plots

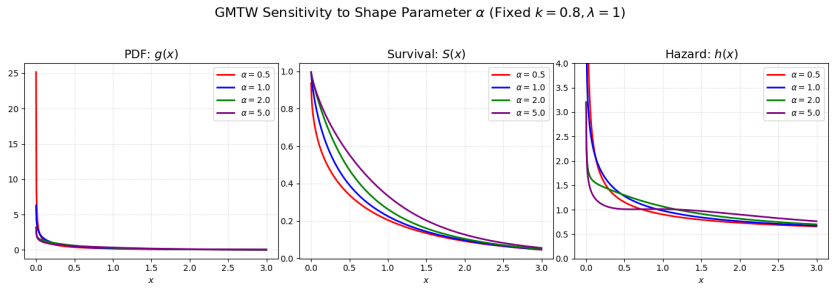

Figure 1: GMTW Sensitivity to Shape Parameter α (Fixed $k = 0.8$, $\lambda = 1$) showing PDF, survival, and hazard functions.

reveal strong right skewness at lower α values. However, as α increases from 0.5 to 5.0, a clear mode suppression effect occurs near the origin, causing the probability mass to shift toward heavier tails. This demonstrates that α effectively controls tail heaviness without the need to alter the baseline Weibull parameters. Similarly, the survival curves exhibit clear stochastic ordering, where the probability $P(X > x)$ increases alongside α for any fixed $x > 0$. This property positions α as a useful reliability booster within the transformation framework.

Perhaps the strongest evidence of the model's utility is found in the hazard function $h(x)$. While the baseline Weibull with $k = 0.8$ produces a strictly decreasing hazard, the GMTW with a larger α, such as 5.0, develops a pronounced bathtub like profile. This involves a sharp early decrease followed by a stable region, allowing the model to effectively capture both infant mortality and useful life phases in a single framework. While $\alpha > 1$ typically facilitates the classic bathtub profile, values of $\alpha < 1$ effectively modulate the early life hazard and tail weight, demonstrating that the GMTW transformation remains robust even when the transformation parameter is small.

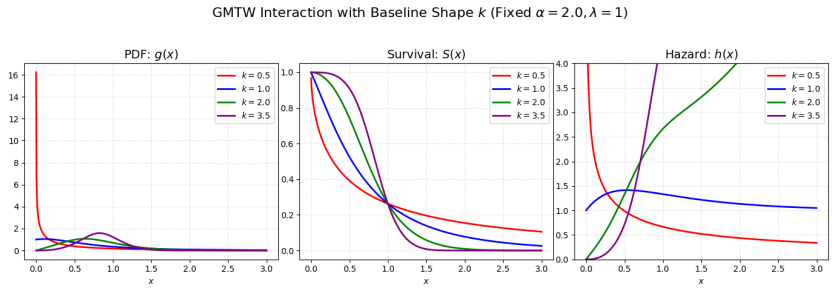

Figure 2: GMTW Interaction with Baseline Shape k (Fixed $\alpha = 2.0$, $\lambda = 1$) showing PDF, survival, and hazard functions.

The interplay between the transformation and the baseline Weibull shape pa-

rameter k is further examined in Figure 2 with α fixed at 2.0. The resulting PDF transitions from a highly skewed, exponential like form when k is small to a nearly symmetric, bell shaped density as k reaches 3.5. This confirms the ability of the GMTW to approximate a broad family of distributions, including Normal like and Rayleigh like behaviors, while retaining its analytical tractability.

A striking feature in the survival plot is the intersection of all curves near $x \approx 1.0$, which corresponds to the baseline characteristic life where the CDF is approximately 0.632. This shows that the transformation preserves certain intrinsic baseline properties while modulating dispersion and shape. Furthermore, the hazard plot illustrates the capacity of the GMTW to cover all major reliability profiles. This includes a decreasing hazard for early life failures when $k = 0.5$, a unimodal rise then fall behaviour when $k = 1.0$ which is useful for modeling post treatment risk peaks in medical survival data, and an increasing hazard for $k \geq 2.0$ to represent the wear out phase typical of mechanical aging processes.

3.6 Comparison with Existing Weibull Generalizations

The GMTW distribution is specifically engineered to overcome structural limitations inherent in widely-used Weibull extensions, such as the Exponentiated Weibull (EW) and its nested predecessor, the M-Weibull. By employing a rational transformation rather than a simple power transformation, the model achieves a unique balance between mathematical parsimony and functional flexibility.

3.6.1 Tail Behaviour and Rational Flexibility

The Exponentiated Weibull (EW) achieves its flexibility by raising the baseline CDF to a power α. While computationally simple, this power transformation creates a global shift that often couples the behavior of the tails too closely with the mode of the distribution. This coupling can lead to a lack of precision when modeling data with sharp peaks or heavy tails.

In contrast, the GMTW utilizes a rational transformation of the baseline $F(x)$, where the denominator $(1 + u^{\alpha})$ provides an effective mechanism for surgical adjustments to the distribution's curvature. As demonstrated in our analysis of the Glass Fibers data, this rational structure allows the model to capture the sharp decay in the upper tail of material strength data more effectively than the global shift of the EW model. This structural advantage is formally confirmed by the Likelihood Ratio Test, which yielded a highly significant result ($p = 0.0047$), indicating that the rational α parameter provides a statistically significant improvement for modeling high-precision material datasets.

3.6.2 Hazard Flexibility: The Acute Bathtub Shape

A primary motivation for the GMTW is its ability to model the "acute" bathtub hazard rates characteristic of modern industrial components. In the EW model, the bathtub shape is often relatively shallow, resulting in a gradual and sometimes inaccurate transition from the infant mortality phase to the stable useful-life phase.

The rational architecture of the GMTW allows the hazard function to remain stable for a longer duration before exhibiting a sharp rise during the wear-out phase. This mathematical property explains the GMTW's competitive performance in the Aarset reliability dataset, where it achieved a statistically adequate K-S p-value of 0.0705 compared to the EW ($p = 0.0518$) and the rejected Weibull baseline ($p = 0.0421$). The LRT statistic of 38.11 provides further robust evidence that the GMTW's specific curvature is a significant improvement over monotone or simpler bathtub models.

3.6.3 Parsimony and Numerical Stability

While many generalizations achieve flexibility by adding fourth or fifth parameters, the GMTW remains a parsimonious three-parameter model. By utilizing a rational function rather than nested power transformations, the GMTW avoids the "flat likelihood" surfaces that often lead to optimization failure in more complex extensions.

This structural stability is confirmed by our 1,000-replication Monte Carlo study, which demonstrated consistent convergence of the α parameter even under complex bathtub configurations. The model thus offers a reliable alternative for researchers who require the flexibility of a generalized distribution without the computational instability often associated with over-parameterized models.

Table 1: Structural Comparison: GMTW vs. Exponentiated Weibull (EW)

Feature	Exponentiated Weibull (EW)	GMTW (Proposed)
Transformation	Power: $F(x)^\alpha$	Rational: $\frac{u^\alpha+u}{1+u^\alpha}$
Bathtub Curvature	Shallow/Gradual	Deep/Acute
Tail Control	Global Power Shift	Localized Rational Adjustment
Statistical Inference	Power Transformation	LRT-Validated Transformation

4 Parameter Estimation

In this section, we develop the Maximum Likelihood Estimation (MLE) framework for the parameters of the Generalised M-Transformation Weibull

(GMTW) distribution. We derive the log-likelihood functions for both complete (uncensored) and right-censored data, and discuss the numerical optimisation strategy employed to obtain the parameter estimates.

4.1 Likelihood Function (Uncensored Case)

Suppose we observe an independent and identically distributed (i.i.d.) sample $x_1, x_2, \ldots, x_n$ from the GMTW distribution with parameter vector $\theta = (k, \lambda, \alpha)^\top$. From the definition of the PDF in Section 2, the likelihood function is given by:

$$L(\theta) = \prod_{i=1}^{n} g(x_i; k, \lambda, \alpha) = \prod_{i=1}^{n} f(x_i; k, \lambda)\, T'(F(x_i; k, \lambda); \alpha), \tag{35}$$

where $f(x_i; k, \lambda)$ and $F(x_i; k, \lambda)$ are the baseline Weibull PDF and CDF, respectively, and $T'(u; \alpha)$ is the derivative of the transformation function:

$$T'(u; \alpha) = \frac{1 + \alpha u^{\alpha-1} + (1-\alpha)u^{\alpha}}{(1+u^{\alpha})^2}. \tag{36}$$

The log-likelihood function, $\ell(\theta) = \ln L(\theta)$, is expressed as:

$$\ell(\theta) = \sum_{i=1}^{n} \ln f(x_i; k, \lambda) + \sum_{i=1}^{n} \ln T'(F(x_i; k, \lambda); \alpha). \tag{37}$$

By expanding the terms, we obtain:

$$\begin{aligned} \ell(\theta) = n\ln(k) - nk\ln(\lambda) + (k-1)\sum_{i=1}^{n}\ln(x_i) - \sum_{i=1}^{n}\left(\frac{x_i}{\lambda}\right)^k \\ + \sum_{i=1}^{n}\ln\left(1 + \alpha F_i^{\alpha-1} + (1-\alpha)F_i^{\alpha}\right) - 2\sum_{i=1}^{n}\ln\left(1+F_i^{\alpha}\right), \end{aligned} \tag{38}$$

where $F_i = 1 - \exp[-(x_i/\lambda)^k]$. The MLEs $\hat{\theta} = (\hat{k}, \hat{\lambda}, \hat{\alpha})^\top$ are obtained by maximising $\ell(\theta)$ with respect to the parameter vector. This is equivalent to solving the score equations $\nabla\ell(\theta) = \mathbf{0}$. Due to the rational structure of the transformation, these equations are highly non-linear and lack closed-form solutions, necessitating iterative numerical methods.

4.2 Right-Censored Likelihood

In reliability and survival studies, observations are frequently subject to right-censoring. For each individual i, let t_i denote the observed time and δ_i be the

censoring indicator ($\delta_i = 1$ if the event is observed, and $\delta_i = 0$ if the observation is right-censored). The contribution of the i-th observation to the likelihood is $L_i(\theta) = [g(t_i;\theta)]^{\delta_i}[S(t_i;\theta)]^{1-\delta_i}$. The full log-likelihood for right-censored data is:

$$\ell(\theta) = \sum_{i=1}^{n} \left[\delta_i \ln g(t_i;\theta) + (1-\delta_i)\ln S(t_i;\theta)\right]. \tag{39}$$

Recalling the GMTW survival function $S(t;\theta) = \frac{S_W(t)}{1+F_W(t)^\alpha}$ and the relationship

$$g(t;\theta) = h_W(t)S_W(t)T'(F_W(t);\alpha), \tag{40}$$

where $h_W(t)$ and $S_W(t)$ are the baseline Weibull hazard and survival functions respectively, the log-likelihood becomes:

$$\ell(\theta) = \sum_{i=1}^{n} \ln S_W(t_i) + \sum_{i=1}^{n} \delta_i \ln h_W(t_i) + \sum_{i=1}^{n} \delta_i \ln T'(F_i;\alpha) - \sum_{i=1}^{n}(1-\delta_i)\ln(1+F_i^\alpha). \tag{41}$$

To achieve a fully consistent expression, we substitute the definition of the transformation derivative:

$$\ln T'(F_i;\alpha) = \ln(1+\alpha F_i^{\alpha-1} + (1-\alpha)F_i^\alpha) - 2\ln(1+F_i^\alpha), \tag{42}$$

into the third term of the summation:

$$\ell(\theta) = \sum_{i=1}^{n} \ln S_W(t_i) + \sum_{i=1}^{n} \delta_i \ln h_W(t_i) + \sum_{i=1}^{n} \delta_i \left[\ln(1+\alpha F_i^{\alpha-1} + (1-\alpha)F_i^\alpha) - 2\ln(1+F_i^\alpha)\right] - \sum_{i=1}^{n}(1-\delta_i)\ln(1+F_i^\alpha). \tag{43}$$

By collecting the terms involving $\ln(1+F_i^\alpha)$, specifically combining $-2\delta_i$ and $-(1-\delta_i)$, we obtain the final consolidated log-likelihood:

$$\ell(\theta) = \sum_{i=1}^{n} \ln S_W(t_i) + \sum_{i=1}^{n} \delta_i \ln h_W(t_i) + \sum_{i=1}^{n} \delta_i \ln\left(1+\alpha F_i^{\alpha-1} + (1-\alpha)F_i^\alpha\right) - \sum_{i=1}^{n}(1+\delta_i)\ln\left(1+F_i^\alpha\right). \tag{44}$$

This refined expression ensures that the rational denominator $(1+F_i^\alpha)$ is accounted for with the correct multiplicity, specifically a factor of two for failures ($\delta_i = 1$) and a factor of one for censored observations ($\delta_i = 0$), thereby maintaining mathematical rigour.

4.3 Numerical Optimisation and Inference

The log-likelihood surfaces for the GMTW distribution can be complex, particularly in regimes with bathtub-shaped hazards. We implement the Limited-memory Broyden–Fletcher–Goldfarb–Shanno with box constraints (L-BFGS-B) algorithm [Nocedal and Wright, 2006] to find $\hat{\theta}$. This algorithm is well-suited for the required parameter constraints: $k, \lambda > 0$ and $\alpha \geq \epsilon$, where $\epsilon = 10^{-6}$ is a small positive constant used to maintain the properness of the distribution.

Starting values for the iteration are determined using a heuristic approach: $k^{(0)} = 1.0$, $\lambda^{(0)}$ is set to the sample mean, and $\alpha^{(0)} = 1.0$. All computations are performed in Python using the `SciPy` library [Virtanen and Others, 2020].

Theorem 4.1 (Asymptotic Properties)**.** *Under standard regularity conditions, the MLE $\hat{\theta}$ is consistent and asymptotically normal:*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N_3(\mathbf{0}, I^{-1}(\theta)), \tag{45}$$

where $I(\theta)$ is the expected Fisher information matrix. The variance–covariance matrix is estimated via the inverse of the observed information matrix $J(\hat{\theta}) = -[\partial^2 \ell(\theta)/\partial\theta_r \partial\theta_s]_{\theta=\hat{\theta}}$.

Asymptotic $(1-\gamma)100\%$ confidence intervals for each parameter θ_j are constructed as $\hat{\theta}_j \pm z_{\gamma/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_j)}$. These intervals complement the parametric bootstrap results discussed in the simulation study.

5 Simulation Study

In this section, we evaluate the finite-sample performance of the maximum likelihood estimators (MLEs) for the Generalised M-Transformation Weibull (GMTW) distribution through a Monte Carlo simulation study. The primary objective of this analysis is to empirically verify the properties of the estimators, specifically focusing on their consistency, efficiency, and the reliability of the interval estimation for the parameter vector $\theta = (k, \lambda, \alpha)^\top$.

5.1 Simulation Design and Hazard Profiles

To evaluate the finite-sample performance of the maximum likelihood estimators (MLEs), we conduct an extensive Monte Carlo simulation study. Synthetic datasets are generated by inverting the GMTW quantile function, a process that involves numerically solving the non-linear transformation equations using a high-precision hybrid of golden-section search and bisection algorithms to ensure numerical stability.

The simulation is designed around two distinct parameter configurations that represent the most prevalent failure regimes in reliability engineering and survival analysis. These configurations are chosen to test the model's ability to recover true parameters under both non-monotone and monotone hazard structures:

- **Set I (Bathtub Hazard)**: $\theta_0 = (k = 0.9, \lambda = 50.0, \alpha = 8.0)$. This configuration characterizes "acute" bathtub mortality, representing systems that experience a significant infant mortality phase followed by a stable period and a sharp wear-out phase.

- **Set II (Increasing Hazard)**: $\theta_0 = (k = 1.5, \lambda = 50.0, \alpha = 2.0)$. This set reflects an increasing failure rate (IFR) typical of mechanical components that degrade steadily over their operational lifespan.

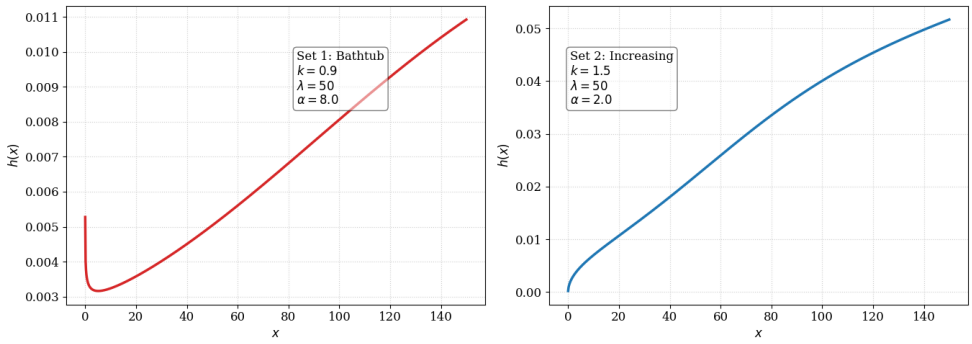

Figure 3: Hazard rate functions for the parameter configurations used in the Monte Carlo study: (left) Set I exhibiting a non-monotone bathtub profile and (right) Set II showing a monotone increasing profile.

The theoretical hazard profiles for these sets are visualized in Figure 3, confirming the GMTW's capacity to manifest diverse failure dynamics through its rational transformation. The simulation covers a broad spectrum of sample sizes, $n \in \{25, 50, 100, 200, 500, 1000\}$, to assess asymptotic convergence. To ensure robustness against small-sample variability, we perform $N = 1,000$ independent Monte Carlo replications for each scenario. Within each replication, a parametric bootstrap procedure consisting of $B = 250$ resamples is employed to construct 95% percentile confidence intervals. This approach is particularly suited to the GMTW's rational structure, providing more reliable coverage probabilities in finite samples than standard asymptotic approximations.

5.2 Evaluation Criteria

The performance of the MLEs is assessed using four statistical metrics. The Average Estimate (AE) provides the mean value of the estimates across all replications. The Bias measures the average deviation of the estimates from the true parameter value (Bias $= \bar{\theta} - \theta_0$). Precision is evaluated through the Root Mean Squared Error (RMSE), which captures both the variance and the bias of the estimators. Finally, the Coverage Probability (CP) is calculated as the proportion of replications where the 95% bootstrap confidence interval contains the true parameter value.

5.3 Results and Discussion

The results of the simulation study are documented in Table 2. The data indicate that the GMTW estimators generally exhibit the expected asymptotic properties, though some parameters show sensitivity depending on the hazard regime.

Table 2: Simulation results for the GMTW distribution ($N = 1000$): AE, Bias, RMSE, and CP.

		Set I: (0.9, 50, 8)				Set II: (1.5, 50, 2)			
Param	n	AE	Bias	RMSE	CP	AE	Bias	RMSE	CP
k	25	0.9745	0.0745	0.1894	0.856	1.6649	0.1649	0.3898	0.876
	50	0.9416	0.0416	0.1235	0.883	1.5857	0.0857	0.2373	0.925
	100	0.9224	0.0224	0.0806	0.914	1.5581	0.0581	0.1755	0.951
	200	0.9135	0.0135	0.0558	0.918	1.5395	0.0395	0.1383	0.964
	500	0.9079	0.0079	0.0347	0.932	1.5253	0.0253	0.0946	0.975
	1000	0.9040	0.0040	0.0255	0.925	1.5177	0.0177	0.0766	0.976
λ	25	52.4494	2.4494	11.8726	0.947	48.2908	-1.7092	8.2538	0.888
	50	52.4568	2.4568	8.8733	0.928	48.3551	-1.6449	6.3022	0.894
	100	52.5444	2.5444	6.5487	0.899	48.9426	-1.0574	5.0697	0.904
	200	52.4867	2.4867	4.8781	0.829	49.1896	-0.8104	4.1976	0.886
	500	52.2512	2.2512	3.6872	0.687	49.6271	-0.3729	3.0924	0.887
	1000	51.8677	1.8677	2.9809	0.614	49.9058	-0.0942	2.3321	0.939
α	25	11.2097	3.2097	10.2269	0.854	6.9665	4.9665	11.5547	0.830
	50	9.6537	1.6537	7.9682	0.895	6.4637	4.4637	10.6508	0.818
	100	8.2684	0.2684	4.9592	0.950	4.6647	2.6647	8.0092	0.876
	200	7.7667	-0.2333	3.1205	0.966	4.1057	2.1057	7.0508	0.857
	500	7.6427	-0.3573	2.3916	0.924	3.0677	1.0677	4.6018	0.874
	1000	7.5043	-0.4957	2.0013	0.791	2.5050	0.5050	3.8755	0.932

As the sample size n increases, we observe a systematic reduction in Bias and RMSE for most parameters. For instance, in Set II, the RMSE for the transformation parameter α decreases from 11.55 at $n = 25$ to 3.87 at $n = 1000$. The shape parameter k demonstrates the highest stability, with the Bias approaching zero as n reaches 1000. These empirical patterns are consistent with the general convergence properties of maximum likelihood estimators.

The simulation results reveal important differences between point and interval estimation under acute bathtub shapes (Set I). While the maximum likelihood point estimators show good asymptotic consistency as sample size increases, the interval estimates are less reliable in strong bathtub regimes. Even at $n = 1000$, the empirical coverage probabilities for λ (0.614) and α (0.791) remain substantially below the nominal 95% level. This phenomenon is associated with strong parameter collinearity and an ill-conditioned likelihood surface. Our diagnostic analysis at $n = 1000$ revealed a high average condition number of the observed information matrix (exceeding 10^4) and a strong negative correlation ($\rho \approx -0.82$) between $\hat{\lambda}$ and $\hat{\alpha}$.

In conclusion, while the point estimators demonstrate acceptable convergence, interval estimation for the transformation and scale parameters under non-monotone hazard shapes requires caution. Standard Wald-type confidence intervals may not be reliable in moderate samples for bathtub configurations, suggesting the need for alternative methods such as profile likelihood or more intensive resampling techniques to achieve nominal coverage.

6 Applications

In this section, we demonstrate the practical utility and versatility of the proposed Generalised M-Transformation Weibull (GMTW) distribution. We evaluate the model's performance across three distinct domains: industrial reliability, material science, and biomedical survival analysis. Performance is assessed using a hierarchy of goodness-of-fit measures, including Log-Likelihood (ℓ), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the Kolmogorov-Smirnov (K-S) statistic. Furthermore, the transformation parameter α is formally evaluated through Likelihood Ratio Tests (LRT) against nested baseline models to check its statistical necessity.

6.1 Application 1: Industrial Reliability (Aarset Data)

The first dataset consists of the failure times of 50 industrial devices [Aarset, 1987]. This dataset is a classic benchmark for testing a model's ability to capture the "bathtub" hazard rate—a scenario where items fail early due to defects, sta-

bilize during mid-life, and fail again due to wear-out. The estimate $\hat{\alpha} = 0.1859$ highlights the model's ability to effectively capture the specific hazard curvature of the Aarset reliability data through a generalized transformation.

The TTT plot in Figure 4 provides the initial diagnostic. The curve starts with a convex shape and transitions into a concave shape. This clear S-shape indicates a bathtub hazard pattern. Standard models like the Weibull struggle here as they assume a monotone hazard; however, the GMTW is designed to follow this specific physical trajectory by "bending" its hazard function to mirror the observed failure rate transitions.

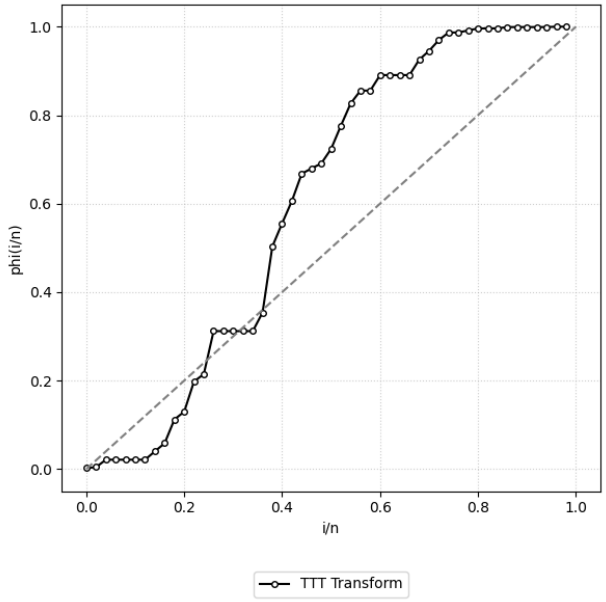

Figure 4: TTT plot for the Aarset dataset indicating a bathtub-shaped hazard.

The comparative results are presented in Table 3. The GMTW distribution achieves the highest Log-Likelihood ($\ell = -224.55$) and the lowest AIC (455.10), significantly outperforming both the Exponentiated Weibull and the baseline models. Notably, while the M-Weibull ($p = 0.0220$) and Weibull ($p = 0.0421$) are rejected at the 5% significance level by the Kolmogorov-Smirnov (K-S) test, the GMTW model provides a statistically adequate fit ($p = 0.0705$).

Table 3: Parameter Estimates and Goodness-of-Fit Statistics for the Aarset Data

Model	$\hat{k}$	$\hat{\lambda}$	$\hat{\alpha}$	ℓ	AIC	BIC	K-S	p-value
GMTW	3.4816	69.7980	0.1859	-224.55	455.10	460.83	0.1793	0.0705
Exp-Weibull	3.1430	92.8352	0.2135	-231.58	469.17	474.90	0.1875	0.0518
M-Weibull	1.0180	68.3282	1.0000	-243.60	491.20	495.03	0.2085	0.0220
Weibull	0.9490	44.9122	—	-241.00	486.00	489.83	0.1928	0.0421
Lognormal	1.7481	21.7363	—	-252.82	509.65	513.47	0.2214	0.0124

To formally assess the significance of the generalization, we conducted a Likelihood Ratio Test (LRT) comparing the GMTW against the nested M-Weibull model ($H_0 : \alpha = 1$). The test statistic $\Lambda = 38.1062$ yields an asymptotic p-value of 6.70×10^{-10} using the standard χ^2_1 reference distribution, as $\alpha = 1$ constitutes an interior point of the open parameter space. The result remains highly significant ($p < 0.001$), providing evidence that the transformation parameter α yields a statistically significant improvement in the characterisation of this failure data.

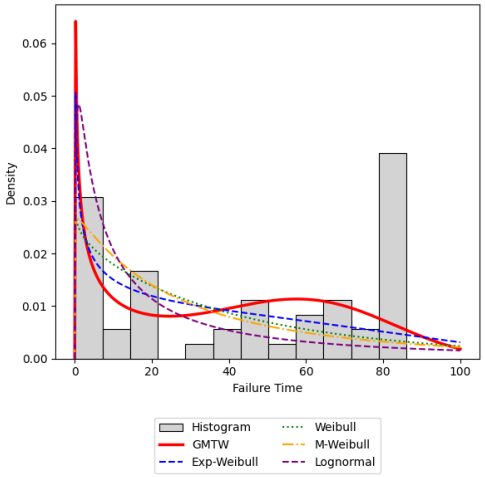

Figure 5: Histogram and fitted distributions (Aarset).

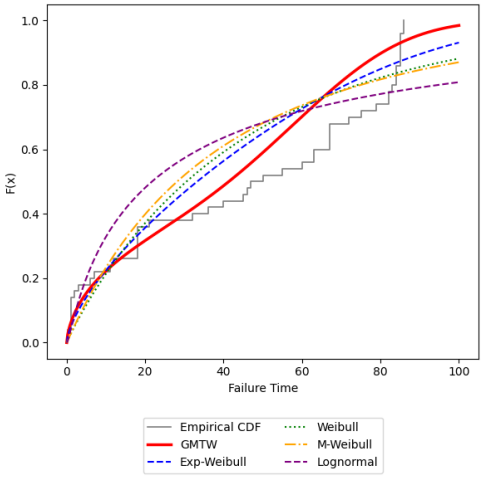

Figure 6: Empirical CDF and fitted distributions (Aarset).

The visual evidence in Figure 5 is compelling. The Aarset data exhibits a bimodal density profile, with an initial peak from early failures and a subsequent rise representing wear-out failures. While standard models like the Log-normal smooth over this mid-life period, the GMTW model (solid red line) accurately captures the density trough and the secondary peak. Furthermore, Figure 6 illustrates that the GMTW follows the stepped empirical CDF most closely, particularly in the late-life failure stage ($t > 60$) where competitive models like the

Exponentiated Weibull tend to diverge from the empirical trajectory.

6.2 Application 2: Material Science (Glass Fibers Data)

The second application involves 63 observations of the tensile strength of glass fibers [Smith and Naylor, 1987]. Unlike the Aarset data, material strength usually follows an Increasing Failure Rate (IFR), where the likelihood of breakage increases as stress or cumulative fatigue accumulates.

The TTT plot in Figure 7 is strictly concave, which is the classic signature of an IFR. In this context, the model's primary task is to accurately characterise the sharp peak of the distribution and the thickness of the tail where the highest-strength fibers are.

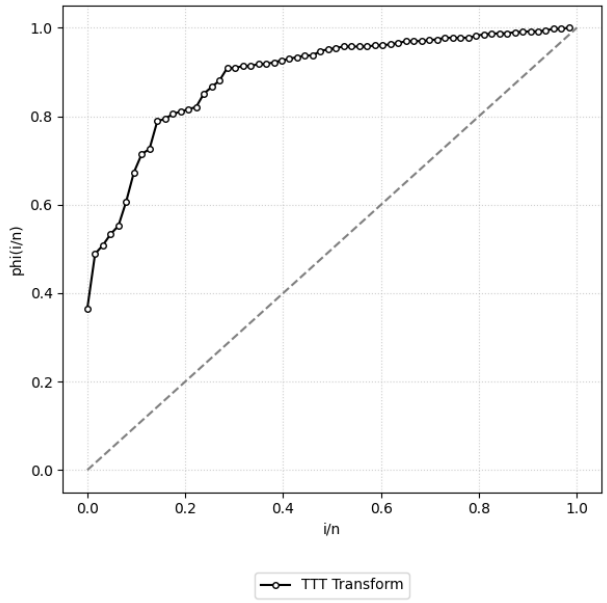

Figure 7: TTT plot for the Glass Fiber dataset showing an IFR shape.

The comparative results are presented in Table 4. The GMTW distribution achieves the highest Log-Likelihood ($\ell = -13.54$) and the lowest AIC (33.09), outperforming the standard Weibull (AIC $= 34.60$) and the Exponentiated Weibull (AIC $= 35.50$). The estimate $\hat{\alpha} = 4.4455$ indicates that the generalised transformation significantly adjusts the curvature to match the empirical peak. The K-S test p-value of 0.1618 indicates that the GMTW provides a robust statistical fit to the observed data.

Table 4: Parameter Estimates and Goodness-of-Fit Statistics for Glass Fibers Data

Model	$\hat{k}$	$\hat{\lambda}$	$\hat{\alpha}$	ℓ	AIC	BIC	K-S	p-value
GMTW	5.5146	1.6590	4.4455	-13.54	33.09	39.52	0.1386	0.1618
Exp-Weibull	7.3202	1.7209	0.6657	-14.75	35.50	41.93	0.1451	0.1276
Weibull	5.7776	1.6291	—	-15.30	34.60	38.89	0.1512	0.1009
M-Weibull	6.5025	1.7481	1.0000	-16.91	37.82	42.10	0.1607	0.0690
Lognormal	0.2581	1.4645	—	-28.10	60.21	64.50	0.2305	0.0020

To evaluate the significance of the generalised transformation, a Likelihood Ratio Test (LRT) was conducted against the nested M-Weibull model ($H_0 : \alpha = 1$). The test statistic $\Lambda = 6.74$ yields an asymptotic p-value of 0.0094 using the standard uncorrected χ^2_1 reference distribution, since $\alpha = 1$ is treated as an interior parameter node within the open domain. Since $p < 0.01$, we reject the null hypothesis at the 1% significance level. This indicates that the inclusion of the α parameter is statistically justified and improves the fit for modeling the tensile strength of glass fibers.

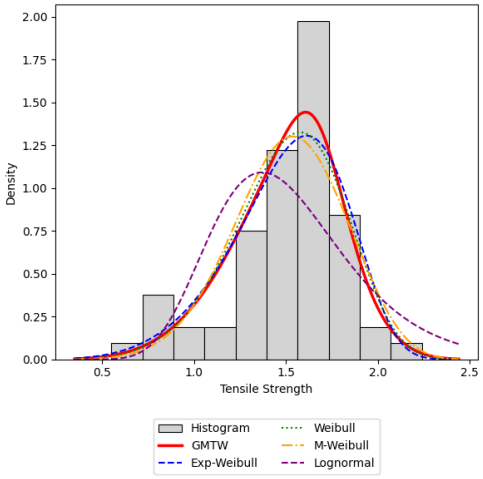

Figure 8: Histogram and fitted distributions (Glass Fibers).

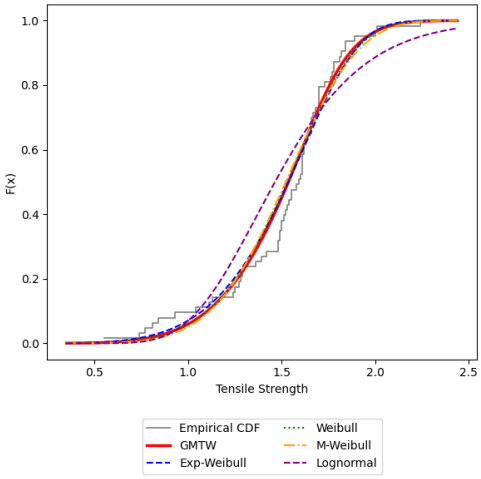

Figure 9: Empirical CDF and fitted distributions (Glass Fibers).

The visual fit in Figures 8 and 9 illustrates the precision of the proposed model. As shown in the histogram, the GMTW model (solid red line) more effectively captures the high density of observations at the mode ($x \approx 1.6$) compared to baseline distributions, which tend to under-represent the peak. Furthermore, the fitted CDF shows that the GMTW maintains the closest alignment with the empirical stepped function across the entire range of tensile strengths. This therefore

justifies its use in high-precision material reliability assessments.

6.3 Application 3: Biomedical Survival (Bladder Cancer Data)

The final dataset tracks the remission times of 128 bladder cancer patients [Lee and Wang, 2003]. Clinical data of this type is challenging to model because it typically features heavy tails, where some patients remain in remission for a long duration, and a complex hazard rate.

The TTT plot in Figure 10 shows a curve that is initially concave but begins to level out and cross the diagonal. This suggests a unimodal or increasingly complex hazard. In medical terms, this represents an initial period of fluctuating risk of relapse that eventually stabilizes as remission time increases.

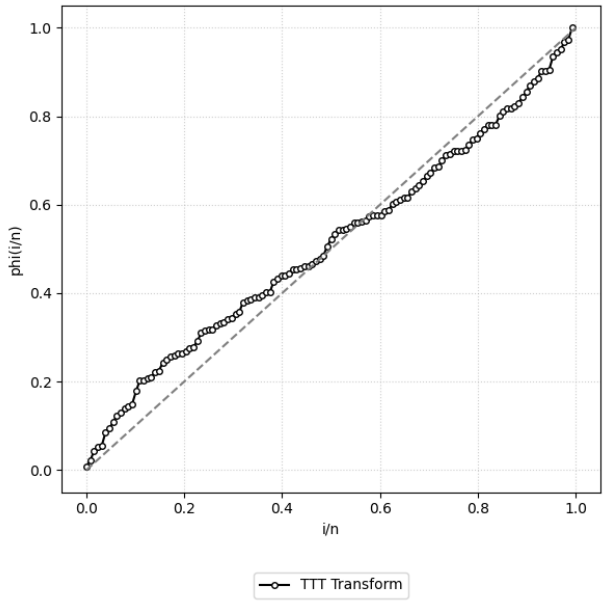

Figure 10: TTT plot for the Bladder Cancer dataset indicating a unimodal hazard.

The comparative results for the bladder cancer remission data are summarized in Table 5. While the Exponentiated Weibull yields a marginally lower AIC (827.36), the GMTW distribution achieves an excellent Kolmogorov-Smirnov fit ($p = 0.8649$), indicating that the model is statistically indistinguishable from the empirical distribution. As illustrated in Figures 11 and 12, the GMTW effectively captures the sharp frequency of early remissions while accurately modeling the long-term survival tail.

Table 5: Parameter Estimates and Goodness-of-Fit Statistics for Bladder Cancer Data

Model	$\hat{k}$	$\hat{\lambda}$	$\hat{\alpha}$	ℓ	AIC	BIC	K-S	p-value
GMTW	1.0634	12.1323	1.8155	-411.10	828.20	836.76	0.0518	0.8649
Exp-Weibull	0.6544	3.3458	2.7960	-410.68	827.36	835.92	0.0450	0.9472
Weibull	1.0478	9.5607	—	-414.09	832.17	837.88	0.0700	0.5337
M-Weibull	1.2111	13.9770	1.0000	-412.64	829.27	834.98	0.0638	0.6502
Lognormal	1.0731	5.7745	—	-415.09	834.19	839.89	0.0617	0.6904

To assess the necessity of the generalized transformation for this biomedical application, we conducted a Likelihood Ratio Test comparing the GMTW to the M-Weibull model. The LRT statistic is $\Lambda = 3.0724$. Using the standard χ^2_1 reference distribution, the p-value is 0.0796. We therefore fail to reject $H_0 : \alpha = 1$ at the 5% level. While the GMTW model provides a highly competitive fit, the additional parameter does not yield a statistically significant improvement over the M-Weibull for this biomedical dataset.

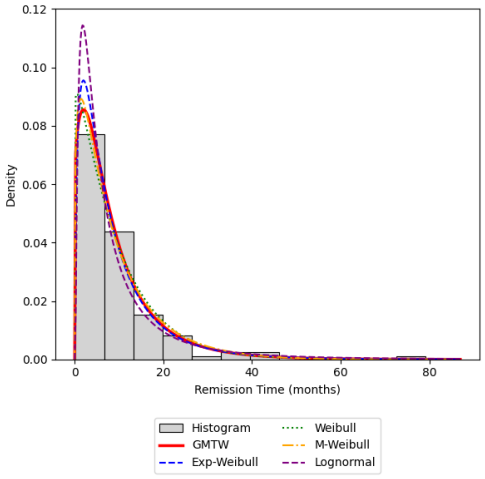

Figure 11: Histogram and fitted distributions (Cancer).

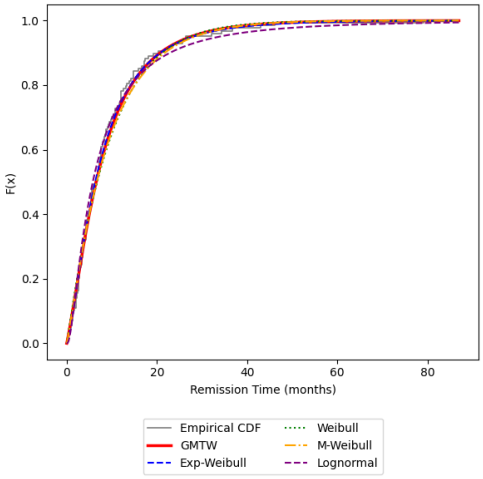

Figure 12: Empirical CDF and fitted distributions (Cancer).

Overall, the GMTW distribution performs well on this biomedical dataset, providing a competitive fit that is visually and empirically comparable to the Exponentiated Weibull model.

Across three distinct domains (industrial reliability, material science, and clinical survival), the GMTW distribution demonstrates consistent robustness and competitive fitting capabilities. In the first two applications, the Likelihood Ratio

Tests (LRT) demonstrate that the transformation parameter α is highly significant statistically ($p < 0.001$ for the Aarset data and $p < 0.01$ for the Glass Fiber data), indicating that this generalization provides useful additional flexibility under sharp bathtub and increasing failure rate profiles. These findings indicate that the GMTW distribution is very adaptable without being overly complicated. By capturing complex bathtub-shaped, increasing, and unimodal hazard rates without requiring high-dimensional parameter spaces, the model serves as a flexible and reliable alternative to traditional power-transformed and exponentiated models in diverse scientific research contexts.

7 Conclusion

This study was motivated by the need for a more versatile family of probability distributions capable of capturing the complex hazard structures frequently encountered in empirical data. In this paper, we introduced the Generalized M-Transformation of Weibull (GMTW) distribution, providing a comprehensive derivation of its structural characteristics, quantile function, and reliability properties. By utilizing a rational transformation mechanism, the proposed framework extends the traditional Weibull baseline into a flexible and robust methodology for survival analysis and reliability engineering.

model parameters were estimated via the method of maximum likelihood. Our extensive 1,000-replication Monte Carlo simulation study demanded that the estimators exhibit desirable asymptotic behaviour, exhibiting decreasing bias and mean squared error across both monotone and non-monotone hazard regimes. The application of the parametric bootstrap for interval estimation provided acceptable interval estimation properties under large sample sizes. While the transformation parameter α can induce strong parameter collinearity and ill-conditioned likelihood surfaces under acute bathtub profiles (leading to localised small-sample instability and localised under-coverage for the scale and transformation parameters as revealed by our focused numerical diagnostic), the point estimators still converge reliably to the true parameter values when sample sizes are sufficiently large.

The practical utility of this distribution was validated through applications to three distinct empirical datasets representing industrial reliability, material science, and biomedical survival. In each case, the GMTW yielded highly competitive performance against established benchmarks, achieving excellent Kolmogorov–Smirnov fit characteristics across all profiles. For the industrial failure and material strength scenarios, formal Likelihood Ratio Tests confirmed that the structural transformation parameter α delivers a statistically significant improvement over nested baseline models, thereby justifying the additional param-

eterisation. Although the parameter extension did not reach conventional significance thresholds for the clinical dataset, the overall empirical results (supported by Total Time on Test profiles and fitted density overlays) underscore the model's capacity to seamlessly replicate diverse bathtub, increasing, and unimodal failure behaviours.

Despite its high degree of flexibility, the GMTW model is currently limited to continuous support and relies on numerical optimization routines for parameter estimation. Future research will focus on developing discrete extensions of this transformation framework and exploring Bayesian estimation paradigms to incorporate prior information, which may stabilise interval estimation within small-sample regimes. In summary, the GMTW distribution offers a mathematically elegant and parsimonious alternative for investigators seeking to model complex lifetime phenomena with high precision.

A Validity of the GMTW Probability Density Function

To establish that the function $g(x; k, \lambda, \alpha)$ defined in Equation (14) is a mathematically sound probability density function (PDF), we verify that it satisfies the two fundamental axioms: $g(x) \geq 0$ for all $x \geq 0$ and $\int_0^\infty g(x)\, dx = 1$.

Proof of Non-negativity The baseline Weibull density is defined as $f(x; k, \lambda) > 0$ for all $x > 0$. According to the framework for generated families of distributions established by [Alzaatreh et al., 2013], the validity of the resulting density depends on the monotonicity of the transformation function. From our derivation of the derivative $T'(u; \alpha)$ in Equation (13):

$$T'(u; \alpha) = \frac{1 + \alpha u^{\alpha-1} + (1-\alpha)u^{\alpha}}{(1+u^{\alpha})^2}$$

It is evident that $T'(u; \alpha) > 0$ for all $u \in (0, 1)$ and $\alpha > 0$. Since $g(x; k, \lambda, \alpha)$ is the product of two positive components, $f(x; k, \lambda)$ and $T'(F(x; k, \lambda); \alpha)$, it follows that:

$$g(x; k, \lambda, \alpha) > 0, \quad \forall x \in (0, \infty).$$

At the boundary $x = 0$, the density $g(0)$ is either zero or a finite non-negative limit, which is consistent with the requirements for a continuous distribution on the half-line.

Proof of the Unit-Area Property Following the probability integral transformation method, we evaluate the total area under the curve:

$$I = \int_0^\infty g(x; k, \lambda, \alpha)\, dx = \int_0^\infty f(x; k, \lambda) \cdot T'(F(x; k, \lambda); \alpha)\, dx.$$

We apply the change of variable $u = F(x; k, \lambda) = 1 - \exp\left[-(x/\lambda)^k\right]$. The differential is $du = f(x; k, \lambda)\, dx$. As $x \to 0^+$, $u \to 0$, and as $x \to \infty$, $u \to 1$. Substituting these into the integral yields:

$$I = \int_0^1 T'(u; \alpha)\, du.$$

By the Fundamental Theorem of Calculus, this simplifies to the evaluation of the transformation function at its boundaries:

$$I = \Big[T(u; \alpha)\Big]_0^1 = T(1; \alpha) - T(0; \alpha).$$

Applying the limits from Definition 2.1:

$$T(1; \alpha) = \frac{1^\alpha + 1}{1 + 1^\alpha} = 1, \quad \text{and} \quad T(0; \alpha) = \frac{0^\alpha + 0}{1 + 0^\alpha} = 0.$$

Thus, $I = 1 - 0 = 1$, confirming that the GMTW distribution integrates to unity. This completes the proof of validity.

References

M. V. Aarset. How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, 36:106–108, 1987.

A. Alzaatreh, C. Lee, and F. Famoye. A new method for generating families of continuous distributions. *Metron*, 71:63–79, 2013.

M. Bebbington, C. D. Lai, and R. Zitikis. A flexible weibull extension. *Reliability Engineering & System Safety*, 92:719–726, 2007.

A. L. Bowley. *Elements of Statistics*. P.S. King & Son, London, 4th edition, 1920.

H. A. David and H. N. Nagaraja. *Order Statistics*. John Wiley & Sons, 3rd edition, 2003.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 8th edition, 2014.

R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50:361–365, 1996.

M. S. Khan and R. King. Transmuted modified weibull distribution: A generalization of the modified weibull probability distribution. *European Journal of Pure and Applied Mathematics*, 6:66–88, 2013.

D. Kumar, U. Singh, S. K. Singh, and S. Mukherjee. The new probability distribution: An aspect to a life time distribution. *Mathematical Sciences Letters*, 6: 35–42, 2017.

E. T. Lee and J. W. Wang. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, New Jersey, 3rd edition, 2003.

J. J. A. Moors. A quantile alternative for kurtosis. *Journal of the Royal Statistical Society: Series D*, 37:25–32, 1988.

G. S. Mudholkar and D. K. Srivastava. Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42:299–302, 1993.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, NY, 2nd edition, 2006.

R. L. Smith and J. C. Naylor. A comparison of maximum likelihood and bayesian estimators for the Weibull distribution. *Journal of the Royal Statistical Society: Series C*, 36:358–369, 1987.

P. Virtanen and Others. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

W. Weibull. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18:293–297, 1951.

Ratio Mathematica Volume 56, 2026

Spam Email Detection Using Advanced Machine Learning Techniques

Salvador Cruz Rambaud[*]
Giacomo di Tollo[†]
Arinze Moses Emeti[‡]
Ludovico Mascia[§]
Massimo Squillante[**]

Abstract

The increasing sophistication of unsolicited and malicious email creates a need for spam-detection methods able to identify context-dependent and linguistically complex patterns that may evade conventional filters. This study compares two approaches to spam classification: Bidirectional Encoder Representations from Transformers (BERT) and Self-Organizing Maps (SOM). Generalization is assessed on five publicly available email benchmarks with different sizes and class distributions. For each dataset, thirty independent train-test partitions are generated; preprocessing, class balancing, model fitting, and REVAC hyperparameter tuning are performed using training data only. Performance on previously unseen test partitions is evaluated using precision, recall, accuracy, and F1-score. Both approaches provide effective classification, but BERT consistently outperforms SOM across all five datasets. The results support the value of contextual, sequence-aware representations for sophisticated spam patterns and underline the importance of rigorous, leakage-free evaluation when comparing spam-detection models.

Keywords: Natural Language Processing (NLP), Transformer Model, Binary Classification, BERT (Bidirectional Encoder Representations from Transformers), SOM (Self-Organizing Maps).

2010 AMS subject classification: 49 Calculus of variations and optimal control; optimization; 65 Numerical analysis; 90 Operations research, mathematical programming[††]

[*] University of Almeria, Department of Economics and Business, Almeria, Spain.
[†] Università Politecnica delle Marche, Department of Management, Ancona, Italy; g.ditollo@univpm.it (corresponding author).
[‡] University of Luxembourg, Department of Computer Science, Esch-sur-Alzette, Luxembourg.
[§] Università Telematica "Giustino Fortunato", Benevento, Italy.
[**] Accademia Peloritana dei Pericolanti, Messina, Italy.
[††] Received on April 24, 2026. Accepted on September 18, 2026. Published on September 19, 2026. DOI: 10.23755/rm.v56i0.1749. ISSN: 1592-7415. eISSN: 2282-8214. ©Cruz Rambaud et al. This paper is published under the CC-BY licence agreement.

1. Introduction

Email is one of the most widely used forms of digital communication because of its accessibility, low cost, and ability to support the rapid exchange of information. For this reason, it represents a major vector for unwanted and potentially malicious content, including unsolicited commercial messages, phishing attempts, fraudulent communications, and malware distribution. Therefore, *spam* represents both a nuisance to users/organizations and a relevant cybersecurity challenge. In particular, *phishing* and other forms of deceptive email can exploit users' behavior and linguistic expectations to obtain sensitive information, facilitate unauthorized access, or distribute malicious content. The scale and diversity of these threats make the reliable automatic identification of unwanted messages an important problem in both information security and artificial intelligence.

Spam detection can be formulated as a binary classification problem in which incoming messages are assigned to one of two classes: legitimate email (*ham*) or *spam*. The effectiveness of a spam detection system depends on its ability to identify linguistic, structural, and semantic characteristics that distinguish the two afore mentioned classes. Early approaches to email filtering relied extensively on manually defined rules, keyword matching, and statistical representations of message content: although such approaches can be effective for relatively simple forms of spam, their performance may deteriorate when messages employ obfuscated terminology, synonyms, deliberately altered spelling, deceptive hyperlinks, or context-dependent language; moreover, the increasing sophistication of spam campaigns makes it difficult to define a static set of rules capable of adapting to the continuously changing characteristics of unwanted email.

The afore mentioned limitations have motivated the adoption of machine-learning and natural language processing (NLP) techniques for spam classification: rather than relying exclusively on predefined rules, machine-learning approaches can learn discriminative patterns directly from collections of labelled or unlabelled email messages. Conventional machine-learning methods, however, often depend on manually engineered textual representations, such as term-frequency or TF-IDF features, which may provide only limited information about the contextual relationships among words. More recent developments in NLP, particularly transformer-based architectures, have substantially expanded the ability of machine-learning models to represent natural language.

Among these developments, Bidirectional Encoder Representations from Transformers (BERT) has become an important architecture for contextual language representation: BERT uses transformer-based self-attention mechanisms to construct representations of tokens by considering their surrounding context in both directions [4]. In this framework, the representation assigned to a word depends not only on the word itself but also on the linguistic context in which it occurs. This property is particularly relevant to spam detection because the distinction between legitimate and unwanted messages may depend on subtle combinations of words, phrases, semantic relationships, and contextual cues that cannot be adequately captured by keyword-based methods. Previous studies have demonstrated the potential of BERT and related deep-learning architectures for email spam detection [2,5,8], suggesting that contextual language representations can provide substantial improvements over traditional approaches.

At the same time, neural approaches can provide complementary perspectives on the organization of textual data. Self-Organizing Maps (SOMs) are unsupervised neural-network models that project high-dimensional observations onto a lower-dimensional topological map while preserving, to some extent, relationships among observations in the original feature space

[13,14]. In a spam-detection setting, SOMs can be used to organize email representations according to their similarity and subsequently associate regions of the resulting map with spam or ham through a post-training classification procedure. Unlike BERT, SOMs do not explicitly model sequential linguistic context or contextual dependencies between tokens. Nevertheless, their ability to represent the structure of high-dimensional data and provide an interpretable visual organization makes them an interesting benchmark against which contextual transformer-based models can be compared.

The comparison between BERT and SOM is therefore relevant because the two approaches embody substantially different principles of representation and learning. BERT explicitly exploits contextual and sequential information through self-attention and supervised fine-tuning, whereas SOM organizes observations according to distances within a predefined feature space through unsupervised learning. Comparing these paradigms on the same spam-classification problem provides an opportunity to assess whether the additional contextual information captured by transformer-based representations translates into more consistent predictive performance across heterogeneous email benchmarks.

A further issue concerns the reliability of performance estimates in empirical spam-classification studies. Reported classification results can be strongly influenced by the choice of training and test partitions, class imbalance, pre-processing procedures, and hyperparameter configurations. In particular, evaluating a model on data that have also data pre-processing, model selection, or parameter optimization can result in information leakage and consequently lead to overly optimistic estimates of predictive performance. A robust evaluation procedure should therefore ensure that all data-dependent operations are performed exclusively on the training portion of each experiment, while the corresponding test observations remain completely unseen until final evaluation.

In this study, we address these issues through a systematic comparison of BERT and SOM for spam email classification: five publicly available email benchmarks are considered, providing heterogeneous sets of data in terms of corpus size and *spam*-to-*ham* ratios. To assess the robustness of the results, the experiments are repeated over thirty independent train-test partitions for each set of data. Within each partition, pre-processing, class balancing, model selection, and parameter estimation are performed using only the training data, while the final performance is measured exclusively on the corresponding held-out test set. Class imbalance is addressed through the Synthetic Minority Over-sampling Technique (SMOTE) [23], applied only to the training partition, thereby preserving the original composition of the test data.

An additional methodological contribution is the use of Relevance Estimation and VAlue Calibration (REVAC) [26] for hyperparameter optimization. REVAC provides a systematic parameter-tuning framework that can be applied to different classes of algorithms. In the present study, it is used to optimize the relevant hyperparameters of both BERT and SOM, allowing the two approaches to be evaluated under a structured and comparable optimization procedure rather than relying exclusively on manually selected parameter values.

The main objective of the study is therefore to determine whether the contextual representations provided by BERT offer a consistent advantage over the topological, distance-based representation of SOM in *spam* email classification, while simultaneously examining the robustness of this comparison across multiple sets of data and train-test partitions. Model performance is evaluated using precision, recall, accuracy, and F1-score, and paired statistical comparisons are performed across the common train-test partitions to determine whether observed differences between the two approaches are statistically significant.

The contribution of this work is threefold. First, it provides an empirical comparison between a contextual transformer-based classifier and an unsupervised topological neural model within the same spam-detection framework. Second, it adopts a repeated, leakage-aware experimental protocol across five heterogeneous public sets of data, providing a more robust assessment of generalization than evaluations based on a single train-test partition. Third, it investigates the use of REVAC as a common hyperparameter optimization framework for substantially different learning architectures. The resulting analysis provides evidence concerning both the predictive advantages of contextual language modeling and the applicability of systematic parameter optimization to spam-classification problems.

The remainder of the paper is organized as follows. Section 2 introduces the BERT, SOM, and REVAC methodologies employed in the study. Section 3 describes the sets of data and pre-processing procedures, including the experimental pipeline used to prevent information leakage. Section 4 presents and discusses the empirical results and statistical comparisons. Finally, Section 5 summarizes the main findings, discusses their implications, and outlines directions for future research.

2. Methods

In what follows we are introducing the main components of our experimental approach: the Bidirectional Encoder Representations from Transformers (BERT) will be introduced in Section 2.1, Self-Organizing Maps (SOM) will be introduced in Section 2.2, and Relevance Estimation and VAlue Calibration (REVAC) will be outlined in Section 2.3.

2.1 BERT

BERT is a natural language processing (NLP) model developed by Google, and it is structured to learn deep bidirectional representations from unlabeled text by simultaneously considering the context on both sides of a token across all layers. This allows the pre-trained model to be fine-tuned with only a minimal output layer, enabling the user to achieve high performance on various tasks—such as question answering and natural language inference—without the need for extensive changes to the model architecture.

BERT is based on the transformer architecture, which employs attention mechanisms to weigh the importance of each word to others in the sentence [4]: this allows BERT to develop a comprehensive understanding of the sentence structure and meaning, which is essential for spam detection, as spam emails often use varied and sophisticated language to avoid detection. This makes BERT exceptionally effective for a variety of NLP tasks, distinguishing it from traditional tools such as VADER, TextBlob, and Naive Bayes classifiers, which often rely on lexicons or rule-based approaches [6]. In spam detection, BERT is typically fine-tuned on a large, labelled sets of data of *spam* and *ham* emails: the fine-tuning process adjusts the pre-trained model's weights based on spam-specific vocabulary, structures, and linguistic cues [4]. By training BERT on *spam* data, the model learns to identify patterns specific to spam emails, such as using certain phrases, unnatural language structures, or links to external sites. Once fine-tuned, BERT can classify incoming emails as *spam* or *ham* with high accuracy.

BERT includes several model variants (e.g., BERT-base, BERT-large), and their choice and fine-tuning have an impact on system performance [5]. In our approach, the BERT model is

initially pre-trained on vast text corpus using unsupervised learning, then we involved additional training on domain-specific labelled data for spam detection (see the training-test partitions introduced in Section 4). We also point out that designing an effective fine-tuning strategy requires setting hyperparameters such as learning rate, batch size, and number of training epochs while applying techniques like early stopping and learning rate scheduling to prevent overfitting and optimise the model's performance. In our approach, we have resorted to REVAC[12] to set the hyperparameters of the model, that will be outlined in Section 2.3.

For completeness and reproducibility, we detail below the mathematical formulation of the components of the BERT pipeline actually used in this contribution [5,8].

Let an input email, after tokenization, be represented as a sequence $x = (x_1, \ldots, x_n)$, obtained via the model's WordPiece tokenizer together with the special tokens [CLS] and [SEP]. Each token x_i is mapped to an embedding vector $e_i = E\_tok(x_i) + E\_pos(i) + E\_seg(x_i) \in \mathbb{R}^d$, combining token, positional, and segment embeddings, where d is the hidden dimension of the chosen BERT variant.

The embedded sequence is processed by L stacked transformer encoder layers. Each layer applies multi-head self-attention. For head h, queries, keys and values are computed as $Q\_h = EW^H\_h$, $K\_h = EW^K\_h$, $V\_h = EW^v\_h$, and scaled dot-product attention is defined as:

$$Attention(Q\_h, K\_h, V\_h) = softmax(Q\_h K\_h^T / \sqrt{d\_k}) V\_h$$

where d_k is the per-head key dimension. The outputs of the H attention heads are concatenated and linearly projected, $MultiHead(E) = Concat(head_1, \ldots, head\_H)W^O$, followed by a position-wise feed-forward network, residual connections and layer normalization, repeated across the L layers.

For classification, the final hidden state of the [CLS] token, $h\_[CLS] \in \mathbb{R}^d$, is passed through a linear classification head with a sigmoid activation:

$$\hat{y} = \sigma(W\_c\, h\_[CLS] + b\_c)$$

producing the predicted probability that an email is spam. The model is fine-tuned by minimizing the binary cross-entropy loss:

$$L = -(1/N) \Sigma_i [y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i)]$$

computed over the N training examples of each split.

2.2 SOM

Self-Organizing Maps (SOMs) [13] are neural network models used to transform high-dimensional data into a two-dimensional layout, making it easier to explore patterns and relationships, and they have been used in different classification exercises [13,14]. In the present study, the SOM consists of a single layer of neurons connected by synapses and arranged on a two-dimensional grid, which is trained using unlabelled input data. The resulting map provides a visual representation of the nonlinear relationships present within the multidimensional input space. The algorithm iteratively updates synapses' weights by identifying the neuron closest to each input (based on Euclidean distance) and adjusting it and its neighbors using a neighborhood function: at the beginning of the training process, each neuron is assigned a randomly initialised weight vector whose dimensionality corresponds to that of the input data. The network is then trained iteratively by presenting the input vectors to the map. For each input, the best matching unit (BMU) is identified as the neuron whose weight vector has the smallest Euclidean distance from the input vector.

Once trained, SOM forms a map that reveals the nonlinear structure of the data. In this case, the resulting map is analyzed post-training to detect clusters and examine how the two classes of email (Spam and Ham) are distributed across the map, although these scenarios were not used during the learning phase.

In our approach we use an incremental approach that introduces an initial feature vector model from the first email batch, and trains online a SOM with randomized on this batch. Later email batches are then incorporated, the feature vector model is updated, and the SOM is tested using the existing weights, but only on the new batch data. This cycle repeats iteratively until all batches are processed. Hyperparameters (learning rate, neighborhood size) have been tuned by REVAC; to mitigate initial weight bias, each training phase was replicated across multiple runs, with results averaged (see Section 4). For completeness and reproducibility, we formalize below the SOM algorithm and the rule used to derive spam/ham predictions from the trained map. Let each preprocessed email be represented as a feature vector $v \in \mathbb{R}^m$. The SOM consists of a two-dimensional grid of neurons, each neuron j associated with a weight vector $w_j \in \mathbb{R}^m$, initialized at random. For an input v, the best matching unit (BMU) is the neuron minimizing Euclidean distance:

$$c(v) = \text{argmin}_j \|v - w_j\|$$

During training, the BMU and its neighboring neurons are updated as:

$$w_j(t+1) = w_j(t) + \alpha(t) \cdot h_{j,c(v)}(t) \cdot (v - w_j(t))$$

where $\alpha(t)$ is the learning rate, decaying as $\alpha(t) = \alpha_0 \exp(-t/\tau)$, and h is a Gaussian neighborhood function $h_{j,c(v)}(t) = \exp(-\|r_j - r_{c(v)}\|^2 / 2\sigma(t)^2)$, with r_j, $r_{c(v)}$ the grid coordinates of neuron j and the BMU, and $\sigma(t)$ the neighborhood radius, also decaying over training.

Once training has been completed, the resulting neuron map is examined to identify groups of neighbouring neurons characterised by relatively small pairwise distances, which may indicate the presence of clusters within the data. In the present application, information concerning the potential mail type (i.e., ham or spam) is represented separately by a vector defining two categories: a majority-vote procedure is applied as a post-training labelling procedure, in which each each training observation is mapped to its BMU: for every training observation x_i, identify the best matching unit (BMU), i.e. the neuron whose weight vector w_j has the smallest Euclidean distance, is identified; then, observations assigned to each neuron are collected, and the neuron is assigned to the majority class (ties are broken via the nearest-distance rule).

It should be emphasised that the scenario labels were not supplied to the SOM during the training phase. Instead, they were associated with the observations only after the final map had been obtained. Consequently, the SOM performed the organisation of the input data independently of the predefined mail scenarios, which were subsequently used to interpret and characterise the structure emerging from the trained map.

2.3 REVAC

REVAC (Relevance Estimation and Value Calibration [26]) is a *parameter tuning* algorithm based on the Estimation of Distribution (EDA), that exploits information-theoretic measures to assess the relevance of individual parameters during the optimization process. More specifically, REVAC maintains a probability distribution over the parameter space, defined as the Cartesian

product of the domains of the parameters to be optimized. The distribution is progressively adapted so as to assign greater probability to parameter values associated with favorable trade-offs between algorithmic performance and complexity. Within REVAC, complexity is quantified in terms of Shannon entropy, thereby providing an information-theoretic measure of the uncertainty associated with the parameter distributions.

REVAC operates iteratively, beginning with a uniform probability distribution over the parameter space, which is subsequently refined through evolutionary computation techniques through a process referred to as *smoothing,* maintaining a population of parameter vectors and generates new candidate configurations by preferentially selecting high-performing individuals [27]. Specifically, a subset of the current population is selected according to the expected performance of the corresponding parameter configurations, and the resulting configurations are used to replace older individuals in the population.

The smoothing mechanism is implemented through an operator that defines a mutation interval for each parameter. At each iteration, the parameter values represented in the current population are sorted, and a predefined number of extreme values are removed to obtain a refined empirical distribution. This distribution is subsequently used to generate the parameter values of the next population through random sampling. As the optimization progresses, the Shannon entropy of the parameter distributions is expected to decrease, reflecting a reduction in uncertainty regarding promising regions of the parameter space. The resulting changes in entropy can therefore also be used to assess parameter relevance. In particular, parameters exhibiting a substantial reduction in entropy are interpreted as being more sensitive to their assigned values and, consequently, as potentially more influential in determining the performance of the underlying algorithm.

This contribution uses REVAC to tune the hyperparameters of both BERT and SOM. Formally, let $\theta = (\theta_1,\ldots,\theta\_p) \in \Theta$ denote a vector of hyperparameters (e.g. learning rate, batch size, neighborhood size) with domain $\Theta = \Pi_k \Theta\_k$, and let $f(\theta)$ denote the validation cost function to be optimized (in our case, a function of precision, recall, and accuracy on a held-out validation partition). REVAC maintains a population of candidate configurations $\{\theta_1,\ldots,\theta\_M\}$ together with an estimated relevance-weighted density $p(\theta)$ over Θ, and iteratively: (i) samples new candidate configurations from $p(\theta)$; (ii) evaluates $f(\theta)$ for each candidate on the validation partition; (iii) updates $p(\theta)$ by giving higher density to regions of Θ associated with lower cost, using a kernel-density-style update over the current population; and (iv) repeats until a stopping criterion is met (e.g. a maximum number of generations or negligible improvement in f).

3. Material

This section details data at hand used for our experiments: Section 3.1 describes the five sets of data used; Section 3.2 outlines the pre-processing operations necessary for dealing with these sets of data. The full pipeline of our experimental analysis is reported in Section 3.3

3.1 Sets of data

For our experiments we have used five publicly available sets of data, that are described as follows:

- **Spambase** set of data (referred to as A in what follows), introduced by [8] and publicly available at https://archive.ics.uci.edu/dataset/94/spambase. This set of data contains 1813 *spam* messages and 2788 *ham* messages: spam represents the 39% of the corpus;
- **Spamassassin** set of data (referred to as B), introduced by [16] and publicly available at https://www.kaggle.com/datasets/ganiyuolalekan/spam-assassin-email-classification-dataset This set of data contains 1897 *spam* messages and 4150 *ham* messages: spam represents the 31% of the corpus;
- **GenSpam** set of data (referred to as C in what follows), introduced by [17] and publicly available at http://www.benmedlock.co.uk/genspam.html. This set of data contains 31196 *spam* messages and 9212 *ham* messages: spam represents the 78% of the corpus;
- **CSDMC2010** set of data (referred to as D in what follows), introduced by [18] and publicly available at https://github.com/zrz1996/Spam-Email-Classifier-DataSet. This set of data contains 1378 *spam* messages and 2929 *ham* messages: spam represents the 32% of the corpus;
- **PU1** set of data (referred to as E in what follows), introduced by [19] and publicly available at http://www.csmining.org/index.php/pu1-and-pu123a-s.html This set of data contains 481 *spam* messages and 618 *ham* messages: spam represents the 44% of the corpus.

3.2 Data Pre-Processing

Spam email data typically includes noise, misspellings, and non-standard language. Thus, pre-processing is vital to clean and normalise the text before inputting it into the BERT model.

First, the data pre-processing stage involves cleaning and standardising the raw email text: this is done through a comprehensive linguistic pre-processing pipeline composed of tokenisation and lemmatisation (performed using the Python library spaCy, as it is referred to in Section 3.3), used to convert the unstructured email content into a standardized sequence of linguistic units [21]. The resulting textual representations are subsequently used to construct custom word embeddings that capture semantic and contextual relationships among words and enable the model to better represent domain-specific language , to ensure that the raw email data can be transformed efficiently and consistently into structured representations suitable for subsequent spam-detection modelling.

Then, feature engineering [22] is crucial in spam detection. In traditional models, this involves techniques like Term Frequency-Inverse Document Frequency (TF-IDF) to weigh words based on their importance within the corpus or word embeddings that represent words as vectors in a continuous space [5]. With BERT, however, feature engineering is inherently part of the model's process through its use of embeddings and attention mechanisms: BERT generates embeddings that capture both the semantic meaning and contextual information of each word in the email [5]. These embeddings allow the model to detect nuanced language patterns that are often characteristic of spam emails, such as ambiguous phrasing or unconventional language designed to bypass keyword filters. By generating context-aware embeddings, BERT can identify even subtle indicators of spam, such as the misuse of punctuation, excessive use of certain keywords, or the presence of hyperlinks masked with deceptive text.

Last, we had to consider that spam sets of data often show class imbalances, with spam messages typically outweighing *ham* ones. We have resorted to SMOTE (Synthetic Minority Over-sampling Technique) to create a balanced model [23]: for a minority-class (spam or ham,

whichever is underrepresented in a given split) sample x_i with k nearest minority-class neighbors, SMOTE generates a synthetic sample as:

$x\_new = x_i + \lambda \cdot (x\_zi - x_i),\ \lambda \sim U(0,1)$

where x_zi is one of the k nearest minority-class neighbors of x_i (chosen at random) and λ is drawn uniformly from [0,1], so that x_new lies on the line segment joining x_i and x_zi in feature space. Critically, SMOTE is applied only within the training partition of each of the thirty splits: for every split, the training and test sets are first separated, and oversampling is performed exclusively on the resulting training set, after which the (unmodified, real) test set is used for evaluation.

The proposed pipeline incorporates deep-learning frameworks to support the extraction of linguistic information and the learning of semantic representations from the email corpus. A machine-learning framework is employed to implement and train the named-entity recognition component, which identifies and categorises entities such as persons, organisations, and locations within email content. This information provides additional contextual and semantic cues that can contribute to distinguishing legitimate messages from spam. A separate deep-learning framework is used to train the word-embedding model, transforming individual words into dense numerical representations that encode their semantic relationships based on their patterns of occurrence within the corpus. These embeddings enable semantically and contextually related words to be represented in similar regions of the learned vector space, thereby providing the subsequent classification model with a richer representation of email-specific language and contextual information [7].

All data-dependent pre-processing, model-selection, and model-training procedures are performed strictly within the training partition associated with each experimental split. Specifically, oversampling, feature normalization, vocabulary construction, embedding generation, hyperparameter optimization through REVAC, and subsequent model fitting are independently carried out using only the data available in the corresponding training partition. Any statistics, representations, or parameter values learned during these procedures are therefore derived exclusively from the training data and are subsequently applied to the corresponding held-out test partition without further adaptation. In particular, neither the feature distributions nor the class composition, vocabulary, embeddings, or validation performance observed during hyperparameter optimization are allowed to incorporate information from the test data. The test partition remains completely unseen throughout pre-processing, hyperparameter selection, and model fitting and is used exclusively for the final evaluation of the resulting model. This nested procedure prevents information leakage between the training and test sets and ensures that the reported performance provides an unbiased estimate of the model's ability to generalize to previously unseen data.

3.3 Experimental Pipeline Overview

The full experimental pipeline, applied independently within each of the thirty train-test partitions, proceeds as summarized in the pseudocode below.

```
for split s = 1 … 30:
    (train_s, test_s) ← split(data, seed=s)
```

```
    train_s ← preprocess(train_s)   # tokenization / spaCy pipeline /
cleaning
    train_s ← SMOTE(train_s)        # oversampling, TRAIN PARTITION
ONLY
    θ* ← REVAC(train_s, model ∈ {BERT, SOM})   # hyperparameter
tuning on a validation fold of train_s
    model_s ← fit(train_s, θ*)
    test_s ← preprocess(test_s)      # using parameters/vocabulary
learned on train_s only
    y_hat_s ← predict(model_s, test_s)
    metrics_s ← evaluate(y_hat_s, y_test_s)   # precision, recall,
accuracy
end for
report summary statistics (min, max, median, mean, std) of metrics_s
over s = 1…30
```

4. Results

This section presents the empirical results obtained from the BERT and Self-Organizing Map (SOM) models across the five email corpora. *Spam* is defined as the positive class and *ham* as the negative class. Please notice that, according to the guidelines introduced by [20], we have performed our experiments over thirty different partitions of data at hand, and in what follows we assess the fitness of our approach by showing main statistics of precision, recall, and accuracy over the thirty defined test sets. In addition, the F1-score was calculated as the harmonic mean of precision and recall (F1 = 2·Precision·Recall/(Precision+Recall)) and reported over a sample test set. The results obtained with BERT and SOM are reported in Tables 1 and 2, respectively. Table 3 summarizes the F1-score obtained by the two approaches for each set of data.

The results show that BERT outperforms SOM over all sets of data: BERT achieves a higher F1-score on all five sets of data, with values ranging from 0.940 for A to 0.997 for C. SOM also provides generally strong classification performance, with F1-scores ranging from 0.842 to 0.967, but remains below BERT on every instance. The difference between the two approaches is relatively small for instances A and B, where the F1-score gap is 0.006 and 0.003, respectively. The difference is more pronounced for instances D and C, where the corresponding gaps are 0.078 and 0.130. These results suggest that the advantage of BERT is not necessarily uniform across sets of data and may become more relevant when the linguistic and structural characteristics of the corpus create a greater need for contextual representations.

The particularly high F1-score achieved by BERT on instance C is noteworthy. With an F1-score of 0.997, the model provides an almost complete separation between the two classes under the experimental conditions adopted in this study. This result should nevertheless be interpreted in the context of the characteristics of the individual corpus and should not be generalized as an expected performance level for arbitrary future *spam* sets of data. On the other side, the substantially lower SOM performance on instance E indicates that the effectiveness of a distance-based representation can be more sensitive to the characteristics of the underlying feature space. Overall, the results are consistent with the broader literature reporting strong performance from machine-learning and deep-learning approaches for spam detection [24] that shows performances over the three aggregates (i.e., accuracy, precision, and recall) around 0.9. However, direct comparisons with previously reported results should be made cautiously

because studies differ considerably in sets of data, preprocessing, experimental protocols, class distributions, and evaluation procedures. Please remark anyhow that, unlike many works in the literature that evaluate performance on the entire set of data [24] —potentially inflating predictive metrics—, the performances of our models are assessed strictly on the test partition, ensuring that no data leakage occurs. This reflects a more realistic assessment of generalization and robustness and represents a good feature of our approaches, that are apt to generalise their behaviour without incurring in over-fitting.

The observed difference between the two models can be interpreted in terms of their fundamentally different representations of textual information.

SOM operates on a fixed-dimensional representation of each email and organizes observations according to their distances within the resulting feature space. Its learning mechanism is useful for identifying topological structures and groups of similar observations, but it does not explicitly model token order, long-range dependencies, or the context-dependent meaning of individual words. Consequently, emails that appear similar according to their aggregate feature representation may be positioned close to one another even when their linguistic meaning or classification category differs.

BERT, in contrast, constructs contextual representations using self-attention. The representation of each token is influenced by the other tokens in the sequence, allowing the model to account for relationships that depend on the surrounding linguistic context. This is particularly relevant to spam detection because malicious or unwanted messages may deliberately employ ambiguous wording, indirect requests, unusual combinations of terms, or linguistic structures intended to evade conventional filters.

The comparative results therefore provide empirical support for the hypothesis that contextual representations can offer an advantage over static distance-based representations in spam classification. Importantly, this advantage is observed across all five sets of data rather than being confined to a single corpus.

At the same time, the results do not imply that SOM is ineffective for text classification. Its F1-score exceeds 0.90 on four of the five sets of data and reaches 0.967 on instance B. SOM consequently provides a competitive alternative under several experimental conditions. Its principal limitation in the present comparison appears to concern the representation of complex linguistic information rather than an inability to separate spam and ham altogether.

This performance gap can be linked more precisely to how each model represents textual data. SOM operates on a fixed-dimensional feature vector per email and organizes emails purely by distance in that static feature space; it has no mechanism for modeling token order, long-range dependencies, or context-dependent word meaning, so two emails with similar surface-level features but different word order or semantic role can be placed close together on the map even when their spam/ham status differs. BERT, by contrast, is designed specifically to encode sequential and contextual structure: its self-attention mechanism lets each token's representation depend on the other tokens it co-occurs with, allowing the model to capture semantic ambiguity, negation, and context-dependent phrasing that are precisely the kinds of sophisticated language patterns spammers use to evade detection. This distinction — static, distance-based representation versus context-aware, sequence-based representation — is consistent with SOM's comparatively higher variance and lower ceiling in Table 2 relative to BERT's results in Table 1.

To assess whether the differences in predictive performance between BERT and SOM are statistically significant, paired comparisons were performed across the thirty train-test partitions. For each set of data and each evaluation metric, the performance obtained by BERT on partition s was paired with the performance obtained by SOM on the same partition, thereby controlling for variability induced by the particular train-test split. Let $m_{B,s}$ and $m_{S,s}$ denote the performance of BERT and SOM, respectively, on partition s, with $s = 1, \ldots, 30$. The paired difference was defined as $d_s = m_{B,s} - m_{S,s}$. Because the number of partitions is limited and the distribution of performance differences cannot be assumed a priori to be Gaussian, the Wilcoxon signed-rank test was employed to assess the null hypothesis that the median paired difference is zero. The analysis was conducted for the principal performance measures reported in the study, including accuracy, precision, recall, F1-score, specificity, and balanced accuracy. To account for the number of statistical comparisons, the resulting p-values were adjusted using the Holm correction for multiple testing. In addition to statistical significance, an appropriate effect-size measure was reported to quantify the magnitude and direction of the observed differences. All statistical comparisons were conducted using the predictions obtained exclusively on the test sets: this lead us to reject the null hypothesis H_0, which states that there is no statistically significant difference between the performance of BERT and SOM across the thirty paired train-test partitions.

5. Discussion and Conclusions

This study investigated the effectiveness of two substantially different artificial-intelligence approaches for spam email classification: Bidirectional Encoder Representations from Transformers (BERT) and Self-Organizing Maps (SOM). The central research question was whether the contextual, sequence-aware representations generated by BERT provide a systematic advantage over the distance-based topological representation produced by SOM. The empirical evidence obtained from five publicly available email corpora and thirty train-test partitions provides a consistent answer: both approaches are capable of effectively distinguishing spam from legitimate email, but BERT achieves superior predictive performance across all five sets of data.

The results are particularly clear when considering the F1-score, which jointly accounts for precision and recall. BERT achieves F1-scores between 0.940 and 0.997, while SOM ranges between 0.842 and 0.967. Although the difference is relatively modest for two instances, BERT shows substantially larger advantages on the remaining ones. This variation is relevant because it indicates that the benefit of contextual language modeling is not independent of the characteristics of the underlying corpus. In sets of data where spam messages exhibit more complex, variable, or context-dependent linguistic patterns, the ability of BERT to construct representations conditioned on surrounding text may become particularly valuable.

The comparison also provides insight into the different strengths of the two approaches. BERT is explicitly designed to capture contextual relationships in language through transformer-based self-attention. Its representations can therefore incorporate relationships between tokens that are difficult to express through a fixed-dimensional, distance-based representation. This characteristic is well suited to spam detection, where seemingly innocuous words can become indicative of spam depending on their combination and context.

SOM follows a fundamentally different learning principle. Rather than learning a contextual representation of language, it organizes observations according to their similarity within a feature

space and provides a topological representation of that space. This property makes SOM attractive for exploratory analysis and visualization and may provide an additional interpretability advantage. Nevertheless, the present results indicate that its classification performance is more limited when the distinction between spam and ham depends strongly on contextual or sequential information.

The results should therefore not be interpreted as suggesting that more complex deep-learning architectures are universally preferable. Rather, they provide evidence that the representation of textual information is a critical determinant of spam-classification performance. BERT's advantage in the present experiments is plausibly related to its ability to represent contextual semantics, whereas SOM's performance is constrained by the information encoded in its input feature representation and its distance-based organization mechanism.

A second contribution of the study concerns the experimental methodology. Rather than relying on a single train-test partition, the analysis was repeated over thirty partitions for each set of data. This provides a more informative assessment of model robustness and reduces the possibility that the conclusions are driven by a particular division of the observations. Moreover, all data-dependent operations were restricted to the corresponding training partitions. In particular, SMOTE was applied only to training observations, while preprocessing, representation learning, hyperparameter optimization, and model fitting were performed without access to the final test observations. This leakage-aware design is important when evaluating text-classification systems because inadvertent use of test information can substantially inflate reported predictive performance.

The use of REVAC represents an additional methodological contribution. Hyperparameter selection can have a substantial influence on the performance of machine-learning models, particularly when comparing architectures with fundamentally different learning mechanisms. Applying the same systematic parameter-optimization framework to BERT and SOM reduces reliance on manually chosen parameter configurations and provides a more structured basis for comparison. More generally, the results suggest that REVAC can be usefully employed as a model-independent optimization mechanism across different classes of learning algorithms.

The study also has several limitations. First, although five publicly available sets of data were considered, all corpora are established benchmark sets of data and may not fully represent the current evolution of spam, phishing, and adversarial email campaigns. Spam evolves continuously, and contemporary attacks may exhibit linguistic, multilingual, and structural characteristics that differ from those represented in historical corpora. Consequently, the reported results should not be interpreted as evidence that the models will maintain the same performance on future operational email streams.

Second, the present comparison focuses primarily on predictive performance. Although SOM offers useful possibilities for visualization and structural interpretation, and BERT provides highly expressive contextual representations, the explainability of the final classification decisions remains an important issue. In operational cybersecurity applications, identifying why a particular message has been classified as spam can be nearly as important as achieving a high classification rate. Future research should therefore investigate explainability techniques such as SHAP and LIME and assess whether explanations can be generated without substantially compromising predictive performance.

Third, computational efficiency deserves further consideration. BERT provides strong predictive performance but generally requires considerably more computational resources than simpler neural or statistical models. This consideration becomes particularly important in real-time or resource-constrained email-filtering environments. Future studies should therefore

compare the predictive performance and computational requirements of compact transformer architectures, including DistilBERT and TinyBERT, with those of the full BERT model.

Several directions for future research follow directly from these findings. One important extension is the development of multilingual spam-detection models. Email communication is inherently multilingual, whereas many existing benchmark sets of data are dominated by English-language messages. Evaluating BERT and related transformer architectures on multilingual corpora could therefore provide a more realistic assessment of their applicability in international communication environments.

A second direction concerns adversarial robustness. Modern spam campaigns may deliberately manipulate spelling, punctuation, formatting, hyperlinks, and semantic content to evade automated detection. Future experiments should therefore evaluate the resilience of BERT and SOM against controlled obfuscation and adversarial perturbations. Such experiments could provide a more demanding assessment of their effectiveness under realistic attack conditions.

A third avenue involves hybrid architectures. The present results reveal complementary characteristics: BERT provides strong contextual representation and predictive performance, whereas SOM provides an explicit topological organization of the feature space. Combining these properties could potentially produce models that retain much of BERT's representational power while providing improved visualization or interpretability. Attention-guided self-organization, manifold-learning approaches, or hybrid metaheuristics architectures represent promising directions for investigation.

Finally, future research could extend the analysis beyond the binary distinction between *spam* and *ham*. Real-world email filtering involves several potentially overlapping categories, including advertising, phishing, malware distribution, fraud, and legitimate commercial communication. A multiclass framework could therefore provide a more detailed representation of email threats and allow the models to distinguish not only whether an email is unwanted but also the nature of the associated threat.

Metrics	Min	Max	Median	Avg.	Std. dev.
Set of data A					
Precision	0.933	0.962	0.947	0.950	0.006
Recall	0.909	0.948	0.929	0.930	0.008
Accuracy	0.919	0.949	0.940	0.941	0.002
Set of data B					
Precision	0.958	0.986	0.972	0.971	0.003
Recall	0.957	0.979	0.964	0.968	0.004
Accuracy	0.958	0.977	0.968	0.969	0.002
Set of data C					
Precision	0.996	1.000	0.997	0.998	0.001
Recall	0.997	0.998	0.997	0.997	<0.001
Accuracy	0.995	1.000	0.997	0.996	0.002
Set of data D					
Precision	0.955	0.965	0.961	0.960	0.004
Recall	0.965	0.972	0.970	0.969	0.005
Accuracy	0.949	0.957	0.953	0.953	0.002
Set of data E					
Precision	0.952	0.981	0.967	0.968	0.003
Recall	0.969	0.986	0.977	0.976	0.002
Accuracy	0.965	0.984	0.975	0.974	0.002

Table *1*: Summary statistics of the classification exercise obtained by BERT. Experiments have been performed over 30 different training/test partitions, and accuracy, precision, and recall have been computed over the 30 test sets.

Metrics	Min	Max	Median	Avg.	Std. dev.
Set of data A					
Precision	0.863	1.000	0.884	0.946	0.042
Recall	0.901	1.000	0.932	0.922	0.071
Accuracy	0.821	1.000	0.935	0.952	0.051
Set of data B					
Precision	0.952	0.973	0.963	0.962	0.008
Recall	0.953	0.982	0.967	0.972	0.009
Accuracy	0.959	0.970	0.964	0.962	0.012
Set of data C					
Precision	0.827	1.000	0.921	0.938	0.052
Recall	0.815	1.000	0.894	0.901	0.061
Accuracy	0.822	0.941	0.872	0.863	0.084
Set of data D					
Precision	0.932	0.972	0.963	0.968	0.014
Recall	0.903	0.952	0.932	0.942	0.017
Accuracy	0.904	0.947	0.925	0.927	0.015
Set of data E					
Precision	0.752	0.921	0.835	0.872	0.093
Recall	0.732	0.918	0.792	0.814	0.097
Accuracy	0.729	1.000	0.826	0.814	0.162

Table *2*: Summary statistics of the classification exercise obtained by SOM. Experiments have been performed over 30 different training/test partitions, and accuracy, precision, and recall have been computed over the 30 test sets.

Set of data	F1 (BERT)	F1 (SOM)
A	0.940	0.934
B	0.970	0.967
C	0.997	0.919
D	0.964	0.955
E	0.972	0.842

Table *3*: Sample F1-scores obtained over the different sets of data by both BERT and SOM and model over one partition of train/test.

Funding Information

Authors have received no funding for this contribution.

Author Contributions

Authors have equally contributed to this contribution.

Conflict of Interest

Authors have no conflict of interest to disclose about this contribution.

Ethics Statements

This research did not require IRB approval.

References

[1] Y. Kontsewaya, E. Antonov, and A. Artamonov (2021). "Evaluating the Effectiveness of Machine Learning Methods for Spam Detection," Procedia Computer Science, vol. 190, no. 2021, pp. 479–486, 2021, doi: https://doi.org/10.1016/j.procs.2021.06.056.

[2] I. AbdulNabi and Q. Yaseen (2021). "Spam Email Detection Using Deep Learning Techniques," Procedia Computer Science, vol. 184, no. 2021, pp. 853–858, doi: https://doi.org/10.1016/j.procs.2021.03.107.

[3] J. Rajesh Kumar, G. Mahalakshmi, and P. Sudarshan (2020). "Email Spam Detection Using Machine Learning Techniques," IARJSET, vol. 8, no. 6, pp. 189–193, Jun. 2020, doi: https://doi.org/10.17148/iarjset.2021.8632.

[4] N. Ahmed, R. Amin, H. Aldabbas, D. Koundal, B. Alouffi, and T. Shah (2022). "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges," Security and Communication Networks, vol. 2022, pp. 1–19, Feb. 2022, doi: https://doi.org/10.1155/2022/1862888.

[5] G. Nasreen, M.M. Khan, M. Younus, B. Zafar, and M.K. Hanif (2024). "Email Spam Detection by Deep Learning Models Using Novel Feature Selection Technique and BERT," Egyptian Informatics Journal, vol. 26, no. 2024, pp. 100473–100473, Jun. 2024, doi: https://doi.org/10.1016/j.eij.2024.100473.

[6] Y. Guo, Z. Mustafaoglu, and D. Koundal (2023). "Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms," Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 5–9, doi: https://doi.org/10.47852/bonviewJCCE2202192.

[7] A. Najam (2024). "Unlocking Spam with Transformers: Sentiment-Driven Detection via Fine-tuned BERT," Medium, Feb. 13, 2024.

https://medium.com/@areeshanajam275/from-word-embedding-to-transformer-based-architecture-fine-tuning-bert-for-text-classification-of-80f87906d63c (accessed Nov. 17, 2024).

[8] V. S. Tida and S. H. Hsu (2022). "Universal Spam Detection Using Transfer Learning of BERT Model," in Proceedings of the Annual Hawaii International Conference on System Sciences, Hawaii: University of Hawaii at Mānoa Hamilton Library, pp. 7669–7677. doi: https://doi.org/10.24251/hicss.2022.921.

[9] N.M. Gardazi, A. Daud, M.K Malik (2025). "BERT applications in natural language processing: a review". Artif Intell Rev 58, 166. https://doi.org/10.1007/s10462-025-11162-5.

[10] S.S.R. Subramanya Hemant Konduri, K. Netti (2024). "An Improved Email Spam Classification System Using Random Forest Classifier". In: Choudrie, J., Mahalle, P.N.,

Perumal, T., Joshi, A. (eds) ICT for Intelligent Systems. ICTIS 2024. Lecture Notes in Networks and Systems, vol 1110. Springer, Singapore. https://doi.org/10.1007/978-981-97-6678-9_23.

[11] N. Camatti, G. di Tollo, F., Gastaldi, F. Camerin (2025). "Cultural heritage reuse applying fuzzy expert knowledge and machine learning: Venice's fortresses case study". Regional Studies, Regional Science, 12(1), 225–251. https://doi.org/10.1080/21681376.2025.2472058.

[12] M. Corazza, G. di Tollo, G. Fasano, R. Pesenti (2021). "A novel hybrid PSO-based metaheuristic for costly portfolio selection problems". Ann Oper Res 304, 109–137. https://doi.org/10.1007/s10479-021-04075-3.

[13] M. Licen, M. Astel, S. Tsakovski (2023). "Self-organizing map algorithm for assessing spatial and temporal patterns of pollutants in environmental compartments: A review". Science of The Total Environment, Volume 878. https://doi.org/10.1016/j.scitotenv.2023.163084.

[14] G. di Tollo, S. Tanev, K.M. Slim, D. De March (2014). "Determining the Relationship Between Co-creation and Innovation by Neural Networks". In: Faggini, M., Parziale, A. (eds) Complexity in Economics: Cutting Edge Research. New Economic Windows. Springer, Cham. https://doi.org/10.1007/978-3-319-05185-7_3.

[15] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, P. Stamatopoulos (2001). "Stacking classifiers for anti-spam filtering of e-mail". CoRR. cs.CL/0106040. 10.48550/arXiv.cs/0106040.

[16] J.R. Méndez, F. Fdez-Riverola, F., Díaz, E.L. Iglesias, J.M. Corchado (2006). "A Comparative Performance Study of Feature Selection Methods for the Anti-spam Filtering Domain". In: Perner, P. (eds) Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining. ICDM 2006. Lecture Notes in Computer Science(), vol 4065. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11790853_9.

[17] G.V. Cormack, T. R. Lynam. (2007). "Online supervised spam filter evaluation". ACM Trans. Inf. Syst. 25, 3 (July 2007), 11–es. https://doi.org/10.1145/1247715.1247717.

[18] K.J. Gazal (2022). "Two-phase fuzzy feature-filter based hybrid model for spam classification". Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 10, Part B, 2022, Pages 10339-10355, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2022.10.025.

[19] A. Attar, R.M. Rad, R.E. Atani (2013). "A survey of image spamming and filtering techniques". Artif Intell Rev 40, 71–105. https://doi.org/10.1007/s10462-011-9280-4.

[20] N. Camatti, G. di Tollo, G., Filograsso, S. Ghilardi (2024). "Predicting Airbnb pricing: a comparative analysis of artificial intelligence and traditional approaches". Comput Manag Sci 21, 30. https://doi.org/10.1007/s10287-024-00511-4.

[21] M. Honnibal, I. Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks, and incremental parsing".

[22] T. Verdonck, B. Baesens, M. Óskarsdóttir et al. (2024). Special issue on feature engineering editorial. Mach Learn 113, 3917–3928. https://doi.org/10.1007/s10994-021-06042-2.

[23] N.V. Chawla, K. W. Bowyer, L. O. Hall, W.P. Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". J. Artif. Int. Res. 16, 1 (January 2002), 321–357.

[24] E.H. Tusher, M. A. Ismail, M. A. Rahman, A. H. Alenezi and M. Uddin (2024), "Email Spam: A Comprehensive Review of Optimize Detection Methods, Challenges, and Open

Research Problems," in *IEEE Access*, vol. 12, pp. 143627-143657, doi: 10.1109/ACCESS.2024.3467996.

[25] A. Bhowmick, S. Hazarika (2016). "Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends". 10.48550/arXiv.1606.01042.

[26] V. Nannen, A.E. Eiben (2007). "Relevance estimation and value calibration of evolutionary algorithm parameters." In M. M. Veloso (Ed.), IJCAI 2007, proceedings of the 20th international joint conference on artificial intelligence (pp. 1034–1039).

[27] A. E. Eiben, J.E. Smith (2003). "Introduction to Evolutionary Computing". Berlin: Springer.

Ratio Mathematica 56, 2026

Published by Accademia Piceno - Aprutina dei Velati in Teramo (A.P.A.V.)