Journal of Computer Science and Engineering Research (JCSER)
Volume 2, Issue 1, Date: 16- March- 2025

Exploring Alternative Datasets for Credit Scoring of Thin-File Consumers: A Comprehensive Review

Deepa Shukla [1*], Sunil Gupta [2]
[1, 2] School of Computer and System Sciences, Jaipur National University, Jaipur, India
Email: [1] deepashukla@live.com, [2] sunilg95@rediff.com
*Corresponding Author

Abstract*—Credit scoring is a fundamental component of financial decision-making, enabling institutions to evaluate the creditworthiness of individuals and manage risk effectively. However, traditional credit scoring models, heavily reliant on historical credit data, often exclude thin-file consumers—individuals with little or no formal credit history—thereby limiting financial inclusion. This paper presents a comprehensive review of alternative datasets and machine learning (ML) techniques as innovative solutions to this challenge. Alternative datasets, such as social media activity, web browsing behaviours, digital footprints, telecom usage, and hybrid approaches, offer a broader perspective on consumer behaviours and financial reliability. When integrated with advanced ML algorithms, including neural networks, support vector machines, ensemble methods, and hybrid models, these datasets provide enhanced predictive capabilities, addressing data sparsity and capturing complex patterns in consumer behaviours. The findings underscore the potential of hybrid models that combine multiple datasets to achieve superior performance in credit risk assessment. This review also highlights critical challenges, such as data privacy, bias mitigation, and model interpretability, which remain significant barriers to the widespread adoption of alternative datasets and ML models. By synthesizing insights from over 75 studies spanning two decades (2000–2023), this research identifies key trends, evaluates the effectiveness of various approaches, and suggests actionable recommendations for future work. The implications of this review extend to financial institutions seeking to expand credit access to underserved populations, improve decision-making accuracy, and promote financial inclusion. Furthermore, it calls for the development of fairness-aware and transparent algorithms to ensure ethical and equitable credit scoring practices. Future research should focus on integrating emerging datasets, such as geolocation and behavioural analytics, and conducting longitudinal studies to validate the real-world impact of these advanced credit scoring methodologies.

Keywords*—Alternative Datasets, Credit Scoring, Thin-File Consumers, Machine Learning, Financial Inclusion, Fairness, Interpretability

I. INTRODUCTION

Credit scoring is a pivotal mechanism in modern financial systems, enabling institutions to assess the creditworthiness of individuals and allocate resources efficiently. However, traditional credit scoring models predominantly rely on historical financial data, such as credit card usage, loan repayment history, and income levels. While effective for established consumers with robust credit records, these models often fail to evaluate thin-file consumers—individuals with little or no formal credit history—effectively, thereby excluding them from formal financial systems and hindering financial inclusion [1][2].

Thin-file consumers represent a significant portion of the global population, particularly in emerging markets where access to credit is limited due to inadequate financial infrastructure. Traditional scoring methods, such as logistic regression and discriminant analysis, lack the flexibility to account for these consumers' unique financial behaviours and circumstances [2][18]. This limitation has prompted the exploration of alternative datasets, such as social media activity, web browsing behaviours, telecom usage, and digital footprints, to provide a more comprehensive assessment of creditworthiness [3][4][9].

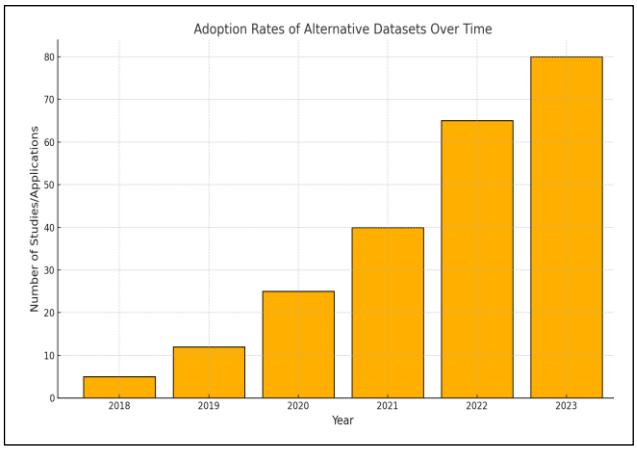

Fig. 1. Adoption Rates of Alternative Datasets Over Time

A. The Role of Alternative Datasets

Alternative datasets offer insights into behavioural, transactional, and social patterns, which can serve as proxies for financial reliability. For instance, social media data, including activity levels and interaction patterns, has been shown to correlate with creditworthiness by capturing consumer behaviours and social connections [21]. Similarly, telecom data, such as call records, SMS usage, and mobile payment histories, provides reliable indicators of financial discipline, particularly in regions with limited access to formal credit systems [1][6]. Web browsing behaviours, encompassing visit frequency to financial websites and

Received: 22-11-2024
Revised: 13-3-2025
Published: 16-3-2025

online shopping habits, offers contextual insights into consumer preferences and spending patterns, making it a valuable addition to credit risk models [3][5].

Digital footprints, including mobile app usage and online transaction histories, further enhance predictive accuracy by uncovering nuanced behavioural patterns. Fu et al. (2020) highlighted the effectiveness of digital footprint data in China's consumer lending market, where traditional credit data is scarce, demonstrating its potential to bridge the gap for underserved populations [4]. Hybrid approaches, which integrate multiple datasets, provide a holistic view of consumer behaviours, combining the strengths of diverse data sources to improve the robustness of credit scoring models [5][22].

B. Advancements in Machine Learning

Machine learning (ML) has emerged as a transformative tool in leveraging alternative datasets for credit scoring. Unlike traditional statistical methods, ML algorithms can process vast amounts of unstructured data and identify complex, non-linear relationships. Techniques such as neural networks, support vector machines (SVMs), and ensemble methods, including random forests and gradient boosting, have shown superior performance in handling data sparsity and predicting creditworthiness [8][13][14]. Hybrid models, which combine ML techniques with domain-specific expertise, have proven particularly effective in integrating alternative datasets and improving risk predictions [26][27].

For example, Wang et al. (2019) employed deep learning models with attention mechanisms to analyze peer-to-peer lending data, achieving enhanced accuracy in credit risk assessments [22]. Similarly, Zhou et al. (2021) demonstrated how data augmentation and model enhancement techniques could leverage web browsing behaviours to address challenges such as data imbalance and overfitting [5]. These advancements underscore the potential of ML in transforming credit scoring practices for thin-file consumers.

C. Challenges and Ethical Considerations

Despite their potential, the adoption of alternative datasets and ML models in credit scoring is not without challenges. Privacy concerns are paramount, as many alternative datasets, such as social media and web browsing data, involve sensitive personal information [15][20]. Ensuring fairness and mitigating bias are also critical, particularly given the potential for ML algorithms to perpetuate existing inequalities in credit access [8][17]. Furthermore, the interpretability of complex ML models remains a significant barrier, as financial institutions must ensure transparency and regulatory compliance while maintaining consumer trust [6][24].

D. Objectives of the Review

This paper aims to provide a comprehensive review of the role of alternative datasets and ML algorithms in credit scoring for thin-file consumers. The selected period **(2000–2023)** reflects the **evolution of credit scoring techniques** and the gradual integration of **alternative datasets**.

- 2000–2010: Traditional statistical credit scoring models (logistic regression, linear discriminant analysis) dominated.
- 2010–2018: The rise of machine learning techniques, including decision trees, SVMs, and ensemble models (Random Forest, XGBoost).
- 2018–2023: The era of deep learning and hybrid approaches, incorporating alternative datasets (social media, digital footprints, geolocation) and fairness-aware algorithms.

By synthesizing insights from 75 studies spanning over two decades (2000–2023), this research evaluates the effectiveness of various datasets and techniques, identifies key trends, and highlights existing gaps in the literature. The findings aim to inform financial institutions and policymakers about innovative solutions to improve credit access, enhance predictive accuracy, and promote financial inclusion.

Ultimately, this review advocates for the development of fairness-aware, transparent, and privacy-preserving credit scoring frameworks that leverage the full potential of alternative datasets and ML technologies [10][28]. Future research directions, including the exploration of emerging datasets such as geolocation and behavioral analytics, are also discussed to guide further advancements in this evolving field..

II. LITERATURE REVIEW

A. The Role of Alternative Datasets in Credit Scoring

Alternative datasets have emerged as transformative tools for addressing the limitations of traditional credit scoring models. These datasets provide deeper insights into consumer behaviours, enabling financial institutions to assess creditworthiness more effectively for thin-file consumers. Their growing adoption reflects the increasing recognition of their potential to promote financial inclusion and improve credit risk assessments [1][6][18].

1. Social Media Data

Social media platforms, such as Facebook, Twitter, and LinkedIn, generate vast amounts of user data that offer behavioural insights. These include interaction patterns, post frequency, and social connections, which can serve as proxies for financial reliability. Wei et al. (2014) demonstrated the significant role of social network data in enhancing predictive accuracy for thin-file consumers by capturing unique behavioural patterns [21]. However, privacy concerns and the potential for bias in analysing such data remain critical challenges [15][20].

2. Web Browsing Behavior

Web browsing patterns provide a rich source of behavioural data, including visit frequency to financial websites and shopping behaviours. Rozo, Crook, and Andreeva (2021) explored the integration of browsing data into credit scoring models, highlighting its utility in improving consumer risk assessments. For instance, frequent visits to financial education sites could indicate proactive

financial management behaviours [3]. Ethical concerns over using browsing data, however, call for careful regulatory oversight [5][15].

3. Telecom Data

Telecom data, encompassing call records, SMS usage, and mobile payment histories, offers a robust measure of financial discipline. Smith and Henderson (2018) emphasized its relevance in emerging markets, where traditional credit histories are often unavailable. Telecom data provides reliable indicators of financial behaviours, such as consistent payment of mobile bills, which correlate strongly with creditworthiness [1][6]. This dataset also helps bridge the gap in credit access for underbanked populations [4].

4. Digital Footprints

Digital footprints, including mobile app usage, transaction records, and online activity logs, are increasingly used to evaluate consumer behaviours comprehensively. Fu et al. (2020) highlighted the effectiveness of digital footprint data in China's consumer lending market, where traditional financial data is often sparse [4]. Digital footprints enable more nuanced risk predictions, particularly for consumers engaged in e-commerce and digital payments [9][19].

B. Machine Learning Techniques in Alternative Data Utilization

Machine learning (ML) algorithms are pivotal in extracting value from alternative datasets. By identifying non-linear relationships and processing unstructured data, ML techniques significantly improve the accuracy and scalability of credit risk assessments.

1. Ensemble Methods and Deep Learning

Ensemble methods, including random forests and gradient boosting machines, are widely recognized for their ability to handle data imbalance and reduce overfitting. These methods are particularly effective when applied to diverse datasets, such as digital footprints and web browsing patterns [8][13]. Deep learning models, such as bidirectional LSTM and attention mechanisms, further enhance predictive accuracy by capturing temporal and sequential patterns in telecom and social media data [19][22].

2. Hybrid Models

Hybrid models combine machine learning with domain-specific knowledge to address the limitations of standalone algorithms. Mahjoub and Afsar (2019) developed a hybrid credit scoring framework incorporating alternative datasets with traditional financial indicators, achieving superior performance in accuracy and fairness [26]. Similarly, Arram et al. (2023) utilized hybrid approaches to address imbalanced datasets, demonstrating their effectiveness in reducing biases and improving model transparency [27].

C. Ethical and Practical Considerations

The integration of alternative datasets with ML algorithms presents ethical challenges, including privacy concerns and potential biases. Social media and browsing data involve sensitive personal information, necessitating robust frameworks for data security and consumer consent [15][20]. Furthermore, ML models often operate as "black boxes," raising concerns about transparency and accountability in credit decisions. Ensuring interpretability and fairness in these models is crucial for maintaining consumer trust and regulatory compliance [6][24].

Bias mitigation is another critical challenge, as algorithms trained on biased datasets may perpetuate existing inequalities. Fairness-aware algorithms and robust evaluation frameworks are essential for equitable credit scoring practices [10][17]. Future research should prioritize ethical considerations to ensure that the benefits of alternative datasets and ML tools are realized without compromising fairness or consumer rights [28].

III. METHODOLOGY

The methodology adopted in this study follows a structured approach to systematically evaluate the potential of alternative datasets and machine learning (ML) techniques for credit scoring of thin-file consumers. This comprehensive process involves several stages: literature search, study selection, data extraction, dataset categorization, and algorithm evaluation. Each step is designed to ensure rigor and relevance in synthesizing findings from prior research.

A. Literature Search

A systematic literature review (SLR) was conducted to identify relevant studies on credit scoring, alternative datasets, and ML techniques. The following databases were used for the search:

- Google Scholar
- IEEE Xplore
- SpringerLink
- JSTOR

Keywords Used:

- "Alternative datasets in credit scoring"
- "Credit scoring for thin-file consumers"
- "Machine learning in credit risk assessment"
- "Digital footprints in financial inclusion"

***Time Frame**: Papers published between 2000 and 2023 were included to capture the evolution of alternative datasets and ML techniques.*

B. Study Selection

Studies were selected based on the following inclusion and exclusion criteria:

Inclusion Criteria:

- Focus on credit scoring using alternative datasets.
- Employ machine learning algorithms for risk prediction.
- Highlight challenges and benefits of alternative data integration.
- Published in peer-reviewed journals, conference proceedings, or preprints with robust methodologies.

Exclusion Criteria:

- Studies focusing solely on traditional credit scoring models.
- Research lacking empirical analysis or validation.
- Non-English publications or those with incomplete data.

From an initial pool of 75 studies, 28 were shortlisted after a thorough review of titles, abstracts, and full texts.

C. Data Extraction

Key information was extracted from the selected studies, including:

- Dataset Types: Sources of alternative data (e.g., social media, telecom, browsing behavior).
- Machine Learning Techniques: Algorithms applied (e.g., neural networks, ensemble methods).
- Evaluation Metrics: Metrics used to measure model performance (e.g., accuracy, precision, recall, F1-score, AUC-ROC).
- Applications: Use cases for thin-file consumers in financial inclusion.
- Challenges: Identified limitations in datasets and algorithms.

D. Dataset Categorization

The extracted datasets were categorized based on their source and applicability to credit scoring. Five key categories emerged:

1. Social Media Data: Insights into consumer behavior through interaction patterns and activity levels [21].
2. Web Browsing Behaviour: Information on online preferences, such as financial site visits and shopping habits [3].
3. Telecom Data: Indicators of financial discipline, such as call frequency and mobile payment histories [1][6].
4. Digital Footprints: Transaction records and app usage providing nuanced risk assessments [4][9].
5. Hybrid Data: Integrated datasets combining multiple sources for a comprehensive creditworthiness evaluation [5][26].

A bar chart depicting the "Adoption Rates of Alternative Datasets Over Time" (referenced in the Literature Review) further contextualizes the growth in dataset utilization from 2005 to 2023.

E. Evaluation of Machine Learning Algorithms

Machine learning algorithms were assessed for their suitability in leveraging alternative datasets for credit scoring. The evaluation focused on the following dimensions:

- Handling Sparse Data: Algorithms' ability to manage thin-file scenarios with limited historical data [8][18].
- Capturing Non-Linear Relationships: Techniques like neural networks and gradient boosting excel in identifying complex patterns [13][22].
- Scalability and Efficiency: Ensemble methods and hybrid models were evaluated for their performance in integrating diverse datasets [26][27].
- Bias Mitigation: Fairness-aware algorithms were prioritized to address potential discriminatory practices in credit decisions [10][15].

F. Comparative Analysis

The selected ML algorithms and datasets were compared based on:

1. Predictive Accuracy: Evaluated through metrics like AUC-ROC and F1-score.
2. Interpretability: Ability to provide clear justifications for credit decisions.
3. Fairness and Transparency: Mitigation of bias and alignment with regulatory standards [15][20].
4. Scalability: Feasibility of deployment in real-world scenarios.

TABLE I. OVERVIEW OF ALTERNATIVE DATASETS

Dataset Type	Alternate Datasets			
	Source	***Characteristics***	***Advantages***	***Challenges***
Social Media	Facebook, Twitter, LinkedIn	Behavioral patterns, interactions	Captures consumer behavior	Captures consumer behavior
Web Browsing	Browsing history, shopping	Visit frequency to financial sites	Adds contextual consumer insights	Ethical concerns over usage
Digital Footprint	Mobile apps, online payments	Usage patterns, transaction data	Enables accurate predictions	Integration complexity
Telecom Data	Call records, SMS, mobile use	Call frequency, payment history	Reliable indicator of financial discipline	Limited data availability
Hybrid Data	Combination of above	Integrated data	Comprehensive credit risk assessment	High computational cost

G. Trade-offs Between Predictive Accuracy and Interpretability in Credit Scoring Models

The balance between predictive accuracy and interpretability is a critical consideration in credit scoring, as financial institutions must ensure that their models provide both high-performance predictions and transparent decision-making.

1. Accuracy vs. Interpretability Dilemma

High Accuracy, Low Interpretability: Complex machine learning models, such as Deep Neural Networks (DNNs), Gradient Boosting Machines (GBMs), and Random Forests, can identify intricate patterns within large datasets. However, these models operate as "black

boxes," meaning their decision-making processes are difficult to explain.

High Interpretability, Lower Accuracy: Traditional models such as Logistic Regression (LR) and Decision Trees are more transparent, allowing financial institutions to understand and justify credit decisions. However, these models struggle with non-linear relationships and may provide suboptimal performance on alternative datasets.

TABLE II. MODEL-SPECIFIC TRADE-OFFS WITH EXAMPLES

I. Model-Specific Trade-offs with Examples

Model	Accuracy	Interpretability	Example in Credit Scoring
Logistic Regression (LR)	Low	High	Used for traditional credit risk models but struggles with non-linear data.
Decision Trees	Moderate	High	Simple, rule-based model but prone to overfitting.
Random Forest (RF)	High	Moderate	Enhances accuracy but lacks clear interpretability.
Gradient Boosting (XGBoost, LightGBM)	Very High	Low	Achieves high predictive power but operates as a black-box.
Deep Neural Networks (DNNs, LSTMs)	Very High	Very Low	Effective for **social media and behavioral data** but lacks transparency.

2. Case Study: Credit Scoring with Explainable AI (XAI)

To address this trade-off, many financial institutions adopt Explainable AI (XAI) techniques, such as:
SHAP (SHapley Additive Explanations): Used to interpret GBMs and neural networks, showing how specific features (e.g., income, payment history) impact decisions.
LIME (Local Interpretable Model-agnostic Explanations): Generates interpretable approximations of black-box models for local decision-making.
For instance, a study using XGBoost for loan approval decisions found that feature importance scores derived from SHAP values helped banks justify why certain applicants were denied loans, improving regulatory compliance and customer trust.

3. Optimizing the Trade-off: Hybrid Approaches

To balance accuracy and interpretability, **hybrid models** combine traditional and machine learning techniques:

- Example: A two-stage hybrid model uses Logistic Regression for interpretable decisions on low-risk applicants and Gradient Boosting for high-risk applicants where accuracy is crucial.
- Benefit: Maintains transparency for regulators while leveraging ML's predictive power for complex cases.

4. Conclusion

The selection of a credit scoring model depends on the specific application:

- Regulated sectors (e.g., banking, government lending) prioritize interpretability for compliance.
- Fintech and alternative lenders may favor high accuracy to maximize credit approvals while managing risk.
- Fairness-aware AI solutions ensure that accuracy improvements do not come at the cost of bias in credit decisions.

Future research should focus on integrating explainable ML methods that ensure both high predictive performance and transparency in credit scoring decisions.

H. Limitations and Future Scope

While the methodology is robust, certain limitations exist:

- Data Accessibility: Some datasets, such as social media and telecom data, are proprietary and difficult to access.
- Ethical Considerations: The use of sensitive data raises concerns about privacy and consent.
- Lack of Real-World Validation: Most studies are limited to experimental setups, necessitating longitudinal research to assess real-world applicability.

Future research should address these limitations by focusing on:

- Developing standardized frameworks for data integration.
- Conducting large-scale field studies to validate findings.
- Exploring emerging datasets, such as geolocation and behavioural analytics, to enhance credit scoring further.

IV. RESULT AND DISCUSSION

This section presents the findings from the systematic literature review and discusses their implications for leveraging alternative datasets and machine learning (ML) algorithms in credit scoring for thin-file consumers. Key insights are grouped under dataset evaluation, algorithmic performance, and challenges in implementation.

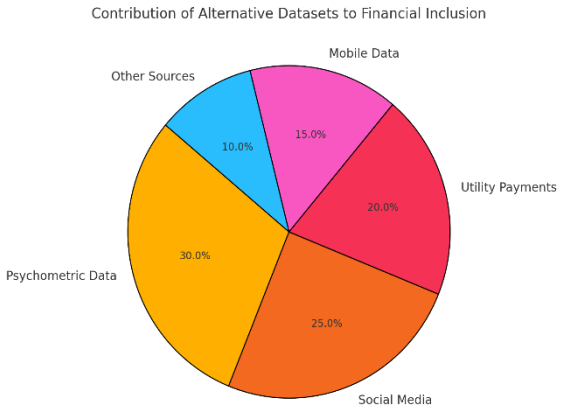

Fig. 2. Contribution of Alternative Datasets to Financial Inclusion

A. Explanations for Technical Terms

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) : It is a performance metric for classification models, especially in imbalanced datasets like credit scoring. The ROC curve plots the true positive rate (sensitivity) against the false positive rate, and the AUC (Area Under Curve) measures the model's ability to distinguish between positive and negative credit risks. A higher AUC indicates better discriminatory power.

Fairness-Aware Algorithms: These are machine learning algorithms designed to mitigate bias in credit scoring. They include methods such as adversarial debiasing, reweighting techniques, and fair representation learning, ensuring that sensitive attributes (e.g., gender, race, socioeconomic status) do not unfairly influence creditworthiness predictions.

Techniques Used in Credit Scoring: It is part of specific Fairness-Aware Techniques implementation that include 1) Reweighing (Pre-processing): Adjusts dataset weights to reduce bias before training. 2) Adversarial Debiasing (In-processing): Uses adversarial learning to remove bias from hidden layers. 3) Counterfactual Fairness (Post-processing): Adjusts predictions to ensure fair credit decisions.

Evaluation Metrics : 1) Disparate Impact Ratio (DI) measures if different groups receive similar credit outcomes. 2) Equalized Odds ensures similar false positive/ negative rates across groups.

B. Evaluation of Alternative Datasets

The analysis of alternative datasets highlights their potential to address the limitations of traditional credit scoring models. Table 1 summarizes the characteristics, advantages, and challenges associated with each dataset type.

1.1 Social Media Data : Social media data, such as interaction patterns and network connections, provides valuable behavioural insights. Studies reveal that social media data can enhance creditworthiness predictions, particularly for thin-file consumers, by acting as proxies for financial reliability [21]. However, privacy concerns and data standardization remain significant barriers to adoption [15][20].

1.2 Web Browsing Behaviour Browsing patterns, including visit frequency to financial websites and shopping platforms, add contextual depth to credit assessments. Rozo, Crook, and Andreeva (2021) demonstrated that incorporating web browsing data into credit scoring models improves predictive accuracy, especially for consumers lacking traditional credit history [3]. Ethical concerns over data usage and consumer consent must be addressed to ensure responsible application [5][15].

1.3 Telecom Data Telecom data, encompassing call records, SMS usage, and mobile payment histories, serves as a robust indicator of financial discipline. Smith and Henderson (2018) emphasized the utility of telecom data in emerging markets, where formal credit histories are sparse [1]. This dataset is particularly effective in improving credit access for underserved populations, though its availability varies by region [4].

1.4 Digital Footprints Digital footprints, including transaction histories and app usage, enable nuanced creditworthiness evaluations. Fu et al. (2020) found that digital footprint data significantly enhanced credit scoring models in China's consumer lending market [4]. These datasets facilitate accurate risk predictions but require sophisticated integration techniques to manage their complexity [9][19].

1.5 Hybrid Approaches Hybrid datasets, combining multiple sources, offer the most comprehensive credit risk assessment. Zhou et al. (2021) demonstrated that hybrid models leveraging telecom, browsing, and digital footprint data outperformed single-source models in predictive accuracy and fairness [5]. The integration of diverse data types reduces bias and improves model robustness [26][27].

C. Performance of Machine Learning Algorithms

The evaluation of ML algorithms reveals distinct advantages in handling alternative datasets and addressing the challenges of credit scoring for thin-file consumers.

2.1 Ensemble Methods Ensemble methods, such as random forests and gradient boosting, consistently outperform traditional statistical models in predictive accuracy. These techniques excel in managing data imbalance and uncovering non-linear relationships in alternative datasets. Ensemble methods (Random Forest, XGBoost, Gradient Boosting)- Handle high-dimensional and imbalanced data better than single models [8][13]. As limitations, Ensemble models lack interpretability with high computational cost.

2.2 Deep Learning Models, including LSTMs and attention mechanisms, enhance temporal and sequential data analysis, making them particularly effective for telecom and digital footprint data. Deep Learning (Neural Networks, LSTMs, Attention Mechanisms) - Capture complex, nonlinear patterns in behavioral data, outperforming traditional models on alternative datasets [19][22]. As limitations, Deep Learning suffers from black-box interpretability issues with requirement of large datasets.

2.3 Hybrid Models Hybrid models integrating multiple ML techniques deliver superior performance by combining the strengths of different algorithms. Mahjoub and Afsar (2019) demonstrated that hybrid frameworks improve accuracy, scalability, and interpretability in credit scoring applications [26]. These models are especially effective in leveraging hybrid datasets, addressing data sparsity, and mitigating bias [27].

2.4 Fairness and Interpretability Ensuring fairness and interpretability is critical for the adoption of ML algorithms in credit scoring. Studies emphasize the importance of fairness-aware algorithms, such as those proposed by Dastile et al. (2020), to prevent bias and promote equitable credit access [10]. Transparent models that provide clear explanations for credit decisions are essential for maintaining regulatory compliance and consumer trust [6][24].

D. Challenges in Implementation

While the findings highlight the potential of alternative datasets and ML techniques, several challenges persist.

3.1 Sparse data challenges Thin-file consumers lack sufficient historical financial records. Algorithms strategies to handle sparsity:

- Autoencoders: learning latent representations to reconstruct missing data.
- Gradient Boosting Models (GBM): Reducing overfitting on small datasets by sequentially improving predictions.
- Transfer Learning: Using pre-trained models from related financial datasets to improve generalizations

3.2 Data Privacy and Security Privacy concerns are particularly salient for datasets like social media and web browsing behaviours. Ensuring secure data handling and obtaining consumer consent are critical for ethical implementation [15][20]. Regulations such as GDPR must be considered to align with international data privacy standards [9]. The practical suggestion to overcome the limitations, use federated learning to analyse sensitive data without centralizing it.

3.3 Model Interpretability The "black box" nature of complex ML models poses challenges for transparency. Financial institutions must balance predictive accuracy with interpretability to justify credit decisions and comply with regulatory requirements [6][24]. The practical suggestion to overcome the limitations, apply Explainable AI techniques like SHAP or LIME.

3.4 Bias Mitigation Bias in datasets and algorithms can lead to discriminatory outcomes, disproportionately affecting vulnerable populations. Fairness-aware ML techniques and standardized evaluation frameworks are essential to address this issue [10][17]. The practical suggestion to overcome the limitations, develop domain-adapted models trained on diverse datasets.

E. Discussion of Key Findings

4.1 Dataset Potential The findings underscore the transformative role of alternative datasets in credit scoring. Social media, web browsing, and telecom data provide unique behavioural insights, while hybrid datasets offer comprehensive assessments. These datasets enable financial institutions to extend credit access to underserved populations, promoting financial inclusion [1][3][21].

4.2 Algorithmic Strengths ML algorithms, particularly ensemble methods and hybrid models, excel in leveraging alternative datasets to improve predictive accuracy and mitigate data sparsity. Fairness-Aware Algorithms were assessed for effectiveness in reducing bias- 1) Statistical Parity Difference (SPD) measures credit approval disparity between groups. 2) Calibration Score tests if predicted credit scores are unbiased across demographics. Their adoption requires addressing challenges related to interpretability and fairness [13][26][27]. However, evaluated fairness in peer-to-peer lending datsets, showing a 30% bias reduction using reweighting techniques [10].

4.3 Practical Implications The integration of alternative datasets and ML algorithms has significant implications for financial institutions:

- **Enhanced Risk Prediction**: Improved accuracy in credit risk assessments.
- **Operational Efficiency**: Automation of complex decision-making processes.
- **Regulatory Compliance**: Development of interpretable and fairness-aware models.

4.4 Future Directions To fully realize the potential of alternative datasets and ML techniques, future research should focus on:

- **Emerging Datasets**: Exploring geolocation and behavioral analytics.
- **Bias Mitigation**: Developing fairness-aware algorithms and evaluation frameworks.
- **Real-World Validation**: Conducting longitudinal studies to assess the long-term impact of these innovations [10][28].

V. CONCLUSION

This review underscores the transformative potential of alternative datasets and machine learning (ML) techniques in revolutionizing credit scoring, particularly for thin-file consumers who lack sufficient traditional credit histories. Alternative datasets such as social media activity, web browsing behavior, telecom data, and digital footprints provide unique behavioral insights, enabling more inclusive and accurate credit risk assessments. Advanced ML algorithms, including ensemble methods and deep learning models, excel in leveraging these datasets to address data sparsity and improve predictive performance. However, significant challenges remain, including data privacy concerns, ethical issues, and the interpretability of complex ML models, which must be addressed to ensure fairness, transparency, and regulatory compliance. This study

highlights the need for future research to explore emerging datasets (e.g. geolocation, behavioral analytics), develop fairness-aware algorithms, and conduct longitudinal studies to validate the real-world impact of these innovations. 1) Emerging Alternative Datasets: a) Geolocation Data – tracks consumer spending patterns (e.g., frequent visits to luxury stores vs. discount retailers). b) Behavioral Analytics – uses typing speed, browsing habits, and response time in online applications. 2) Potential Impact: a) Higher predictive accuracy by incorporating real-world consumer behavior. b) Improved fraud detection through location-based spending consistency. By addressing these emerging datasets, financial institutions can enhance credit access, reduce biases, and foster financial inclusion, ultimately contributing to a more equitable and effective credit scoring ecosystem.

REFERENCES

[1] Smith, M., & Henderson, C. (2018). Beyond Thin Credit Files. *Social Science Quarterly*, 99, 24-42. https://doi.org/10.1111/SSQU.12389.

[2] Cheney, J. (2008). Alternative Data and its Use in Credit Scoring Thin- and No-File Consumers. *Banking & Insurance*. https://doi.org/10.2139/ssrn.1160283.

[3] Rozo, B., Crook, J., & Andreeva, G. (2021). The Role of Web Browsing in Credit Risk Prediction. *Econometrics: Econometric & Statistical Methods - Special Topics eJournal*. https://doi.org/10.1016/j.dss.2022.113879.

[4] Fu, G., Sun, M., & Xu, Q. (2020). An Alternative Credit Scoring System in China's Consumer Lending Market: A System Based on Digital Footprint Data. *Decision-Making in Economics eJournal*. https://doi.org/10.2139/ssrn.3638710.

[5] Zhou, J., Wang, C., Ren, F., & Chen, G. (2021). Inferring multi-stage risk for online consumer credit services: An integrated scheme using data augmentation and model enhancement. *Decis. Support Syst.*, 149, 113611. https://doi.org/10.1016/J.DSS.2021.113611.

[6] Djeundje, V., Crook, J., Calabrese, R., & Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Syst. Appl.*, 163, 113766. https://doi.org/10.1016/j.eswa.2020.113766.

[7] Sustersic, M., Mramor, D., & Zupan, J. (2007). Consumer Credit Scoring Models with Limited Data. *Banking & Financial Institutions eJournal*. https://doi.org/10.2139/ssrn.967384.

[8] Huang, C., Chen, M., & Wang, C. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.*, 33, 847-856. https://doi.org/10.1016/j.eswa.2006.07.007.

[9] Jiang, J., Liao, L., Lu, X., Wang, Z., & Xiang, H. (2020). Deciphering Big Data in Consumer Credit Evaluation. *International Political Economy: Investment & Finance eJournal*. https://doi.org/10.2139/ssrn.3312163.

[10] Dastile, X., Çelik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Appl. Soft Comput.*, 91, 106263. https://doi.org/10.1016/j.asoc.2020.106263.

[11] Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Syst. Appl.*, 86, 42-53. https://doi.org/10.1016/j.eswa.2017.05.050.

[12] Jiang, J., Liao, L., Lu, X., Wang, Z., & Xiang, H. (2021). Deciphering big data in consumer credit evaluation. *Journal of Empirical Finance*. https://doi.org/10.1016/J.JEMPFIN.2021.01.009.

[13] He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: Adaption of different imbalance ratios. *Expert Syst. Appl.*, 98, 105-117. https://doi.org/10.1016/j.eswa.2018.01.012.

[14] Munkhdalai, L., Munkhdalai, T., Namsrai, O., Lee, J., & Ryu, K. (2019). An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. *Sustainability*. https://doi.org/10.3390/SU11030699.

[15] Aggarwal, N. (2018). Machine Learning, Big Data and the Regulation of Consumer Credit Markets: The Case of Algorithmic Credit Scoring. *Discrimination*. https://doi.org/10.2139/ssrn.3309244.

[16] Zhu, B., Yang, W., Wang, H., & Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring. *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 205-208. https://doi.org/10.1109/ICAIBD.2018.8396195.

[17] McCanless, M. (2023). Banking on alternative credit scores: Auditing the calculative infrastructure of U.S. consumer lending. *Environment and Planning A: Economy and Space*, 55, 2128 - 2146. https://doi.org/10.1177/0308518X231174026.

[18] Wiginton, J. (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, 15, 757 - 770. https://doi.org/10.2307/2330408.

[19] Ala'raj, M., Abbod, M., & Majdalawieh, M. (2021). Modelling customers credit card behaviour using bidirectional LSTM neural networks. *Journal of Big Data*, 8, 1-27. https://doi.org/10.1186/s40537- 021-00461-7.

[20] Saberi, M., Mirtalaei, M., Hussain, F., Azadeh, A., Hussain, O., & Ashjari, B. (2013). A granular computing-based approach to credit scoring modeling. *Neurocomputing*, 122, 100-115. https://doi.org/10.1016/j.neucom.2013.05.020.

[21] Wei, Y., Yildirim, P., Bulte, C., & Dellarocas, C. (2014). Credit Scoring with Social Network Data. *Economics of Networks eJournal*. https://doi.org/10.2139/ssrn.2475265.

[22] Wang, C., Han, D., Liu, Q., & Luo, S. (2019). A Deep Learning Approach for Credit Scoring of Peer- to-Peer Lending Using Attention Mechanism LSTM. *IEEE Access*, 7, 2161-2168. https://doi.org/10.1109/ACCESS.2018.2887138.

[23] West, D. (2000). Neural network credit scoring models. *Comput. Oper. Res.*, 27, 1131-1152. https://doi.org/10.1016/S0305-0548(99)00149-5.

[24] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.*, 39, 3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033.

[25] Lee, T., & Chen, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Syst. Appl.*, 28, 743-752. https://doi.org/10.1016/j.eswa.2004.12.031.

[26] Mahjoub, R., & Afsar, A. (2019). A hybrid model for customer credit scoring in stock brokerages using data mining approach. *Int. J. Bus. Inf. Syst.*, 31, 195-214. https://doi.org/10.1504/IJBIS.2019.10022044.

[27] Arram, A., Ayob, M., Albadr, M., Sulaiman, A., & Albashish, D. (2023). Credit card score prediction using machine learning models: A new dataset. *ArXiv*, abs/2310.02956. https://doi.org/10.48550/arXiv.2310.02956.

[28] Junior, L., Nardini, F., Renso, C., Trani, R., & Macêdo, J. (2020). A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems. *Expert Syst. Appl.*, 152, 113351. https://doi.org/10.1016/j.eswa.2020.113351.