AI-Based Fatality Risk Hotspot and Cause-Shift Prediction System Using Accidental and Natural Death Data

Vaibhav M [1], Sathvik M V [2], Vamshi Krishna P S [3], Varun S[4], Dr. Savitha G[5]

[1] *Student, Department of Computer Science Engineering R V Institute of Technology and Management, Bengaluru, India*

[2] *Student, Department of Computer Science Engineering R V Institute of Technology and Management, Bengaluru, India*

[3] *Student, Department of Computer Science Engineering R V Institute of Technology and Management, Bengaluru, India*

[4] *Student, Department of Computer Science Engineering R V Institute of Technology and Management, Bengaluru, India*

[5] *Associate Professor, Department of Computer Science Engineering R V Institute of Technology and Management, Bengaluru– 560076, Karnataka, India*

Abstract - *The increasing number of accidental and natural deaths reported each year highlights the need for intelligent systems that can support early risk detection and preventive planning. Although large volumes of mortality data are collected by government agencies, these datasets are commonly used only for historical reporting and statistical summaries. This paper proposes an AI-Based Fatality Risk Hotspot and Cause-Shift Prediction System that transforms historical death records into predictive insights. The proposed framework applies machine learning techniques to estimate future fatality counts, classify dominant causes of death, and identify high-risk regions through clustering. Random Forest Regression is used for trend forecasting, Random Forest Classification is applied for cause prediction, and K-Means Clustering is used to group regions based on fatality severity. A web-based dashboard is designed to present results through interactive charts and region-wise analysis. The system enables authorities to move from reactive reporting to proactive decision-making by supporting public safety planning, healthcare preparedness, and targeted interventions. Experimental observations indicate that the integrated approach improves interpretability and provides useful predictions for policy development and resource allocation.*

Key Words*: *Artificial Intelligence, Fatality Prediction, Machine Learning, Risk Hotspots, Cause Analysis, Public Safety, Clustering

1.INTRODUCTION

Accidental and natural deaths remain a serious public concern, causing significant social and economic impact every year across different regions of India. Large volumes of mortality data are collected by government agencies and statistical departments, covering factors such as state, year, and cause of death [1], [5]. However, these records are often used only for annual reports and historical summaries, which limits their value for future planning. Traditional analytical methods mainly describe past trends and are less effective in predicting upcoming risks, identifying vulnerable regions, or detecting shifts in fatality patterns [3]. This creates a strong need for intelligent, data-driven, and cost-effective approaches that can support proactive decision-making.

Artificial Intelligence, which combines computational models with data analysis techniques, has emerged as a powerful solution for extracting meaningful insights from structured datasets [5]. By leveraging machine learning algorithms and historical mortality records, AI systems can uncover hidden patterns, forecast future fatalities, and classify dominant causes of death with higher efficiency than manual methods [2]. These capabilities allow administrators to recognize risk hotspots, allocate healthcare resources effectively, and plan preventive measures before losses increase.

This paper aims to examine the role of AI and machine learning in transforming mortality analysis into a predictive framework. The following sections discuss the use of regression models for fatality forecasting, classification techniques for cause-shift detection, clustering methods for hotspot identification, and interactive dashboards for supporting policy decisions and public safety management [4].

2. LITERATURE REVIEW

The application of Artificial Intelligence in public data analysis has significantly changed the way large-scale datasets are interpreted and utilized. Instead of relying only on manual reports and descriptive statistics, modern systems use predictive models to generate future insights from historical records [1]. In mortality analysis, AI enables authorities to shift from reactive observation to proactive planning by forecasting risks, identifying vulnerable regions, and detecting changes in death patterns. The major areas of research include fatality trend prediction, hotspot identification, cause classification, and intelligent decision-support systems.

2.1 Fatality Trend Prediction:

Forecasting future fatality counts is one of the most important tasks in mortality analytics. Traditional statistical methods such as linear regression and time-series models were initially used to estimate yearly death trends. While these methods perform adequately on simple datasets, they often struggle to capture non-linear relationships and interactions between multiple variables such as region, year, and cause category [2].

Machine learning models, especially ensemble methods such as Random Forest Regression, have shown improved performance in trend forecasting because they can learn complex patterns from historical data. These models are robust to noise, reduce overfitting, and provide more reliable predictions across diverse regions [3]. In healthcare and population studies, regression-based forecasting has been successfully applied to estimate disease burden, accident frequency, and resource demand

2.2 Hotspot Identification Using Clustering:

Another important area of research focuses on identifying high-risk locations where fatalities are consistently higher than average. Clustering algorithms are widely used for this purpose because they automatically group regions with similar fatality characteristics without requiring predefined labels [4].

K-Means Clustering is one of the most commonly used techniques due to its simplicity and efficiency. By analyzing death counts and related features, the algorithm can divide regions into categories such as low-risk, medium-risk, and high-risk zones. Such grouping helps administrators prioritize interventions and allocate emergency resources where they are most needed [5]. Similar clustering approaches have also been used in crime mapping, epidemic surveillance, and disaster management systems.

2.3 Cause Classification and Cause-Shift Analysis:

Several studies have explored classification techniques for determining dominant categories within public datasets. In mortality analysis, classification models can predict whether accidental or natural causes are more likely to dominate in a specific region or future period. This is valuable because the preventive actions required for accidents differ significantly from those needed for health-related natural deaths [6].

Algorithms such as Decision Trees, Support Vector Machines, and Random Forest Classification are frequently used for category prediction. Among them, Random Forest performs effectively on structured datasets because it combines multiple decision trees and improves classification accuracy. Cause-shift analysis extends this idea by identifying transitions in dominant causes over time, helping authorities detect emerging risks early [7].

2.4 Computational Drug Repositioning:

Recent research emphasizes the importance of integrating predictive models with visualization tools. Standalone predictions are less useful if decision-makers cannot interpret results quickly. Therefore, web dashboards and interactive systems are increasingly combined with AI models to present charts, regional comparisons, and real-time recommendations [8].

Frameworks such as Streamlit and Flask have enabled the development of lightweight decision-support platforms where users can select a region, enter future years, and instantly view predictions. These systems improve accessibility for non-technical users and make AI outputs more practical for governance and planning

2.5 Research Gap:

Although previous studies have addressed forecasting, clustering, or classification separately, limited work combines all three capabilities in one unified framework for accidental and natural death data. Most existing systems focus only on historical reporting or a single analytical method. This paper addresses that gap by proposing an integrated AI system that combines fatality prediction, hotspot detection, cause-shift classification, and dashboard-based visualization in a single platform [9].

3. OBJECTIVES

The primary objective of this research is to develop an intelligent analytical system that can transform historical mortality records into predictive insights for better planning and decision-making. The specific objectives of the proposed work are as follows:

- To analyze accidental and natural death data collected across different regions and years.
- To predict future fatality counts using machine learning regression techniques.
- To identify high-risk regions by applying clustering algorithms on mortality patterns.
- To classify dominant causes of death as accidental or natural using supervised learning models.
- To detect shifts in fatality causes over time for early intervention and policy planning.

- To design a web-based dashboard for interactive visualization of predictions and trends.
- To support government agencies, healthcare departments, and safety authorities with data-driven recommendations.

4. DISCUSSION

4.1 Challenges & Limitations

Despite the growing importance of AI in mortality prediction and hotspot analysis, several limitations still affect the reliability and practical performance of these systems. These challenges arise in key areas such as fatality forecasting, cause classification, clustering accuracy, and data integration, often leading to reduced prediction quality and higher implementation complexity.

In fatality forecasting, regression models may produce inaccurate estimates when historical data contains sudden anomalies, incomplete records, or irregular reporting patterns [13]. The prediction process is generally expressed as $Y = f(X)$, where Xincludes variables such as region, year, and previous fatalities, while Yrepresents the expected future death count. If the training data does not properly capture changing real-world conditions, the estimated value of Ymay deviate significantly from actual outcomes.

Similarly, classification models used for cause-shift prediction may face errors when accidental and natural death patterns overlap significantly. If the training dataset is imbalanced or contains non-representative samples, the classifier may become biased toward majority classes and perform poorly on minority patterns [14], [20]. This can lead to incorrect predictions regarding whether accidental or natural causes will dominate in future periods.

Hotspot detection using clustering techniques also presents challenges. Algorithms such as K-Means depend strongly on the selected number of clusters and the initial centroid positions. The clustering objective is to minimize the total within-cluster variance using the function $J = \sum_{i=1}^{k} \sum_{x \in C_i} || x - \mu_i ||^2$, where $C_i$denotes a cluster and $\mu_i$represents its centroid. Poor initialization or unsuitable values of $k$may lead to unstable grouping results or incorrect separation of high-risk and low-risk regions [15].

Another major limitation is data integration. Mortality datasets collected from multiple sources may contain inconsistent labels, missing values, duplicate entries, and varying reporting standards. These issues require extensive preprocessing before model training and may reduce the dependability of the final predictions [15].

Overall, these challenges highlight the need for balanced datasets, robust preprocessing, continuous model validation, and the inclusion of external factors such as demographics or environmental conditions. Addressing these limitations is essential for transforming AI-based mortality systems into reliable tools for real-world decision-making.

4.2 Emerging Role of AI and ML

Artificial Intelligence (AI) and Machine Learning (ML) are becoming powerful solutions for many limitations found in traditional mortality analysis systems, such as forecasting errors, biased classification results, weak hotspot detection, and difficulties in handling incomplete or heterogeneous datasets [13]. Conventional statistical methods often struggle with large and dynamic datasets, whereas AI models can process complex relationships and generate more accurate predictions from historical records.

Particularly in fatality forecasting and cause-shift prediction, AI techniques improve speed, scalability, and predictive performance by learning hidden patterns across multiple years and regions. Advanced learning models can analyze structured mortality data, detect subtle trends, and adapt to changing patterns more effectively than rule-based systems [16], [17].

By using ensemble learning and multitask frameworks, AI models can simultaneously predict future death counts, classify dominant causes, and estimate risk levels while capturing relationships between different variables. This improves model robustness on imbalanced datasets and reduces errors caused by noisy or incomplete records [14], [19]. Higher predictive accuracy enables early identification of high-risk regions and supports timely policy interventions.

Automated Machine Learning (AutoML) further simplifies model development by selecting algorithms, tuning hyperparameters, and optimizing performance with minimal manual effort. This increases accessibility for non-expert users and improves the efficiency of deploying predictive systems for public administration [14], [19]. Deep learning approaches can also enhance pattern recognition when larger datasets become available, especially for long-term trend forecasting and anomaly detection.

Overall, AI and ML not only address the weaknesses of traditional analytical methods but also create opportunities for integrated intelligent systems that combine forecasting, classification, clustering, and real-time visualization. Such advancements can significantly strengthen data-driven governance, healthcare preparedness, and public safety planning [16], [17].

4.3 Impact of Personalized and Region-Specific Decision Support

By enabling highly customized planning strategies, AI-driven mortality systems have the potential to transform public administration from a one-size-fits-all model into a region-specific decision framework [18]. Instead of applying the same preventive measures everywhere, authorities can use localized predictions to design actions based on the risk profile of each region.

By combining multiple data sources such as mortality records, demographic indicators, healthcare availability, and environmental conditions, intelligent systems can forecast region-specific fatality risks, identify vulnerable populations, and optimize intervention strategies. This can improve efficiency while reducing unnecessary expenditure and delayed responses [15], [18].

For example, urban areas with rising accidental deaths may require stronger traffic monitoring, industrial safety audits, or emergency response infrastructure. Regions showing increasing natural deaths may benefit from improved hospitals, disease surveillance, and awareness programs. Such targeted planning ensures that resources are allocated where they create the greatest impact.

In the future, these systems can be expanded with AI-driven recommendations that automatically suggest policy actions based on predicted risks. Integration with real-time data streams, GIS mapping, and population analytics can further enhance responsiveness and improve administrative outcomes [18].

Overall, the broader impact includes cost reduction through smarter resource allocation, better preparedness through early warnings, and improved governance through evidence-based planning. With responsible data use and continuous model refinement, AI-based mortality analytics can become a cornerstone of safer and more efficient public systems in the coming years [18].

4.4 Applications and Practical Significance

The proposed AI-Based Fatality Risk Hotspot and Cause-Shift Prediction System has wide practical applications in governance, healthcare, and public safety. By transforming historical mortality data into predictive insights, the system helps institutions make informed decisions rather than relying only on past reports.

One major application is in government policy planning. Authorities can identify regions with rising fatality trends and introduce preventive regulations, awareness campaigns, or infrastructure improvements before the situation worsens [21]. Predictive analytics can also support budget planning by directing resources toward high-priority districts.

In public safety management, the system can be used to monitor accident-prone areas and implement corrective measures such as traffic control, industrial inspections, and emergency preparedness programs. If certain regions consistently appear as high-risk hotspots, targeted interventions can be carried out to reduce future losses [22].

Another important application is healthcare resource allocation. Forecasts of natural deaths can help hospitals and health departments prepare beds, staff, medicines, and emergency facilities in advance. Early planning improves response efficiency and reduces pressure during peak demand periods [23].

The proposed model can also contribute to smart city administration by integrating with digital dashboards, GIS platforms, and urban planning systems. Risk maps and future forecasts can guide safer city design, road management, and disaster readiness strategies [24].

Overall, the system serves as a practical decision-support tool that converts raw data into meaningful actions for society.

4.5 Future Scope

Although the current system demonstrates the usefulness of AI in mortality analytics, several enhancements can further improve its real-world value. One important future direction is the integration of real-time data sources such as hospital records, traffic systems, weather feeds, and emergency response data. Real-time updates would make predictions more dynamic and timely [25].

The use of deep learning models can also be explored for capturing highly complex temporal patterns and large-scale datasets. Techniques such as Long Short-Term Memory (LSTM) networks and hybrid neural models may improve long-term forecasting accuracy [26].

Another valuable enhancement is the addition of GIS-based heatmaps and geospatial intelligence. Visual hotspot maps can help administrators instantly identify vulnerable areas and monitor regional changes more effectively [27].

Future systems may also include population, economic, and environmental indicators to create richer predictive models. Factors such as population density, healthcare access, pollution levels, and climate conditions can improve risk estimation accuracy [28].

Finally, deploying the platform as a mobile and cloud-based application can increase accessibility for field officers, disaster management teams, and local authorities. With continuous improvement, the proposed framework can evolve into a national-scale intelligent safety monitoring system.

4.6 Performance Evaluation and Model Validation

To ensure that the proposed system is reliable for real-world use, proper model evaluation is essential. Different machine learning tasks require different performance metrics. For fatality count prediction, regression models are commonly assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). These metrics measure how closely predicted values match actual death counts. Lower MAE and RMSE values indicate better forecasting accuracy, while a higher R^2value reflects stronger explanatory power [29].

For cause-shift classification, performance can be evaluated using Accuracy, Precision, Recall, and F1-Score. Accuracy measures the overall correctness of predictions, while Precision indicates how many predicted positive cases were correct. Recall evaluates how effectively the model identifies relevant cases, and F1-Score balances both Precision and Recall. These metrics are especially useful when accidental and natural death classes are unevenly distributed [30].

Clustering models are often validated using measures such as the Silhouette Score and inertia values. The Silhouette Score indicates how well data points fit within their assigned clusters compared to other clusters. A higher score suggests clearer separation between low-risk, medium-risk, and high-risk regions. Inertia measures the compactness of clusters and is minimized during K-Means optimization [31].

Cross-validation techniques can further improve trustworthiness by testing the model on multiple train-test splits instead of a single dataset partition. This reduces the risk of overfitting and ensures that performance remains stable across different samples [32].

4.7 Ethical and Social Considerations

While AI-based mortality prediction offers many benefits, ethical considerations must also be addressed. Public datasets may contain sensitive demographic or health-related information, making data privacy an important concern. Proper anonymization, secure storage, and controlled access mechanisms are necessary before deploying such systems [33].

Another issue is algorithmic bias. If historical data reflects reporting gaps or unequal representation of regions, the model may produce unfair predictions that disadvantage certain communities. Regular auditing, balanced datasets, and transparent evaluation processes are required to reduce such risks [34].

The system should also be viewed as a decision-support tool rather than a replacement for human judgment. Final policy decisions must involve domain experts, healthcare officials, and administrators who can interpret predictions within real-world contexts. Responsible use of AI ensures that technology supports fairness, accountability, and public welfare.

5. CONCLUSIONS

The increasing availability of mortality data creates a valuable opportunity to improve public safety and administrative planning through intelligent analytics. However, traditional reporting systems mainly describe past events and provide limited support for forecasting future risks. This research addresses that gap by proposing an AI-Based Fatality Risk Hotspot and Cause-Shift Prediction System using accidental and natural death data.

The proposed framework combines multiple machine learning techniques to perform different analytical tasks within a single platform. Random Forest Regression is used to estimate future fatality counts, Random Forest Classification is applied to predict dominant causes of death, and K-Means Clustering is used to identify low-risk and high-risk regions. By integrating these methods with an interactive dashboard, the system transforms raw historical records into meaningful and accessible insights.

The study demonstrates that AI can significantly improve the usefulness of mortality datasets by enabling proactive rather than reactive decision-making. Predictive insights can help governments strengthen safety measures, improve healthcare readiness, allocate resources efficiently, and plan targeted interventions for vulnerable regions.

Although challenges such as data quality, imbalance, and changing external conditions still exist, continuous model improvement and richer datasets can further enhance system reliability. Future integration with real-time data, geospatial tools, and advanced deep learning models can make such systems even more effective.

In conclusion, the proposed work highlights how Artificial Intelligence can play an important role in transforming mortality analysis into a smart decision-support framework. With responsible implementation and ongoing refinement, AI-based predictive systems can contribute to safer communities, stronger governance, and better long-term planning..

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[2] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, vol. 1, pp. 281–297, 1967.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011.

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[5] World Health Organization, "WHO Mortality Database: Global Health Estimates and Mortality Trends," Geneva, Switzerland, 2024.

[6] Government of India, Ministry of Home Affairs, "Annual Statistical Report on Accidental and Natural Deaths in India," New Delhi, India, 2024.

[7] IndiaAI, "Accidental and Natural Death Statistics Dataset Repository," Government of India Digital Platform, 2025.

[8] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.

[10] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[11] S. Raschka and V. Mirjalili, Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[13] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

[14] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[15] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 2nd ed. Boston, MA, USA: Pearson, 2018.

[16] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, Sept. 1999, doi: 10.1145/331499.331504.

[17] N. Bhatia and V. Dixit, "Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis," International Journal of Computer Applications, vol. 79, no. 2, pp. 15–22, Oct. 2013.

[18] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Canada, pp. 1137–1143, 1995.

[19] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[20] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[23] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.

[24] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[25] J. Brownlee, Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End. Melbourne, Australia: Machine Learning Mastery, 2020.

[26] Ministry of Road Transport and Highways, Government of India, "Road Accidents in India – Annual Report," New Delhi, India, 2024.

[27] National Crime Records Bureau, "Accidental Deaths and Suicides in India," Ministry of Home Affairs, Government of India, New Delhi, India, 2024.

[28] Ministry of Health and Family Welfare, Government of India, "National Health Statistics Report," New Delhi, India, 2024.

[29] Smart Cities Mission, Government of India, "Urban Safety, Risk Monitoring and Smart Planning Report," New Delhi, India, 2024.

[30] World Bank, "Population Density, Development Indicators and Urban Risk Statistics," Washington, DC, USA, 2024.

[31] European Commission, "Ethics Guidelines for Trustworthy Artificial Intelligence," Brussels, Belgium, 2021.

[32] OECD, "Recommendation of the Council on Artificial Intelligence," Paris, France, 2023.

[33] United Nations, "Global Population and Mortality Trends Report," Department of Economic and Social Affairs, New York, NY, USA, 2024.

[34] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 5th ed. Hoboken, NJ, USA: Wiley, 2012.

[35] J. Fox, Applied Regression Analysis and Generalized Linear Models, 3rd ed. Thousand Oaks, CA, USA: Sage Publications, 2015.

[36] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[37] A. Ng, Machine Learning Yearning: Technical Strategy for AI Engineers, DeepLearning.AI, 2018.

[38] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[39] R. Szeliski, Computer Vision: Algorithms and Applications, 2nd ed. Cham, Switzerland: Springer, 2022.

[40] M. Wooldridge, An Introduction to MultiAgent Systems, 3rd ed. Hoboken, NJ, USA: Wiley, 2020.

[41] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[42] A. Géron, "Practical Deep Learning for Structured Data Systems," O'Reilly Media Technical Series, Sebastopol, CA, USA, 2023.

[43] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, 3rd ed. Cambridge, U.K.: Cambridge University Press, 2020.

[44] G. Shmueli, N. R. Patel, and P. C. Bruce, Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python. Hoboken, NJ, USA: Wiley, 2017.