A Hybrid Tree-Based Neural Architecture for Driver Attrition Prediction in Ride-Hailing Platforms

B. Naresh[1], K. Srinivas[2*], J. Vijayalakshmi[3], K. Rajkumar[3], Kandala Sahitya[3], Mohd Saifuddin[3]

[1]Assistant Professor, [2]Associate Professor & Head, [3]UG Student, [1,2,3]Department of Computer Science and Engineering (Data science)

[1,2,3]Vaagdevi Engineering College, Bollikunta, Warangal, 506005, Telangana, India
*Correspondence: K. Srinivas (srinivas_k@vecw.edu.in)

Abstract

Driver attrition has emerged as a critical concern for ride-hailing platforms, as it directly influences operational efficiency, service continuity, and customer satisfaction. Identifying drivers who are at high risk of leaving the organization at an early stage enables companies to implement targeted retention strategies and maintain workforce stability. Conventional approaches to attrition analysis primarily rely on manual surveys, basic statistical techniques, and rule-based systems. However, these methods depend heavily on historical insights and human judgment, limiting their ability to effectively process large-scale, high-dimensional data and uncover complex, hidden relationships among influencing factors. To address these limitations, the proposed system introduces a hybrid deep learning and machine learning framework for accurate driver attrition prediction. A Long Short-Term Memory (LSTM) neural network is employed to learn and extract meaningful latent representations from structured driver data, capturing intricate feature dependencies and underlying patterns. These extracted features are subsequently utilized by a Greedy Tree (GT) classifier to perform the final classification, ensuring both improved predictive performance and model interpretability. For benchmarking purposes, traditional machine learning models, including Random Forest (RF), Gradient Boosting (GB), and Support Vector Classifier (SVC), are implemented and evaluated. Experimental results demonstrate that the proposed Greedy Long Short-Term Memory Tree (GLSTMT) model achieves superior performance, attaining an accuracy of 100% on the dataset used in this study. The complete system is deployed using a Flask-based web architecture, facilitating seamless model integration, secure user authentication, and efficient interaction between frontend and backend components.

Keywords: Long Short-Term Memory (LSTM), Greedy Tree Classifier, Hybrid Machine Learning Model, Deep Learning, Workforce Retention.

1. Introduction

Driver attrition refers to the natural process in which drivers leave an organization due to various reasons, such as resignation, career changes, or dissatisfaction with job conditions. It has become a significant concern for organizations, particularly in the ride-hailing and transportation sectors, where workforce stability directly impacts operational performance. Attrition occurs when drivers leave the organization at a faster rate than they are hired, leading to unfilled vacancies and reduced productivity. A high attrition rate indicates frequent employee turnover, which can result in increased recruitment costs, loss of experienced personnel, and disruption of organizational efficiency, as shown in figure. 1. Therefore, controlling and managing attrition is essential for sustaining organizational growth and stability. Driver attrition can be categorized into several types, which help in understanding the underlying causes. Voluntary attrition occurs when drivers choose to leave the organization on their own, while involuntary attrition happens when employment is terminated by the organization. External attrition refers to drivers leaving for opportunities in other organizations, whereas internal attrition involves role changes or promotions within the same organization. Measuring the attrition rate is crucial for identifying patterns and addressing key factors contributing to employee turnover. It is typically calculated by dividing the number of drivers who leave the organization by the average number of drivers over a specific period.

Impact of components in driving employee retention	Average economic context	Economic downturn context	Rank difference between contexts
1 (Most Impactful Factor)	Opportunity	Work	+1 More Important
2	Work	People	+1 More Important
3	People	Opportunity	-2 Less Important
4	Rewards	Rewards	No Change in Importance
5 (Least Impactful Factor)	Organization	Organization	No Change in Importance

Figure. 1: Driving Employee Attrition Strategy

Recent statistics highlight the severity of the issue. Studies indicate that nearly one-third of new drivers leave within the first six months of employment. According to the Job Openings and Labor Turnover Survey (JOLTS), approximately 3 to 4.5 million employees leave their jobs each month in the United States. Reports from the Bureau of Labor Statistics show attrition rates reaching as high as 57.3% in certain sectors, while the average across industries is around 19%. Additionally, the Society for Human Resource Management (SHRM) estimates the average cost per hire to be approximately USD 4,129. A retention rate of 90% is generally considered healthy, implying that organizations should aim to maintain attrition rates below 10% for sustainable growth.

2. Literature Survey

Denget al. [1] explored how to optimize the operations of RHPs by investigating the impact of commission rates on drivers' switching behaviours in a dynamic mobility market. Two queue-theory-based mathematical models were developed to examine the relationship between commission rates, drivers' switching behaviours, and critical platform parameters in optimizing the operations of RHPs. Numerical examples were presented to demonstrate the applicability of these models in determining the optimal commission rate to enhance the operations of RHPs under duopoly and fully competitive market conditions. Cohen and Zhang [2], developed an endogenous model to simulate the operations of two-sided RHPs, leading to the determination of the optimal service price and the most appropriate commission rate in an analytical manner.

Sun and Ertz [3] propose a system dynamic modelling framework to explore the ride-hailing operations, discovering that commission rates, order prices, and investment levels are the critical variables for optimizing the operations of RHPs. Zhong et al. [4] examine how RHPs compete with traditional taxi companies, resulting in designing specific pricing strategies under unregulated and regulated environments in optimizing the operations of RHPs. The studies above have provided insights into how to improve the operations of RHPs from different perspectives. Sun et al. [5] discovered two types of order assignment strategies have distinct impacts on the operations of RHPs.

Xu et al. [6] proposed a generalized fluid framework incorporating appropriate commission rate setting for modelling ride-hailing to improve the operations of RHPs. Ke and Qian [7] developed an optimal pricing scheme for improving the operations of RHPs. Bandiera et al. [8] present a mathematical model to simulate the interaction between service providers and customers. Xi et al. [9] developed a multi-leader–multi-follower model in which platforms compete to optimize their operations.

Yao and Zhang [10] adopt a many-to-many matching framework for solving the pricing problem that RHPs face in a multi-modal transport network. Sun et al. [11] investigated the influence of various regulatory policies on the operations of RHPs with respect to the growth of individual platforms. The discussion above has provided a better understanding of how to improve the operations of RHPs from three perspectives. There is, however, a lack of studies on how to achieve a subtle trade-off between platform profitability and customer service quality in optimizing the operations of RHPs. The interplay between commission rates, drivers' switching behaviours, and critical platform parameters in optimizing the operations of RHPs is unclear. There are some limitations in this study. First, there is specific simplification about drivers' behaviours in developing the two mathematical models for optimizing the operations of RHPs in the dynamic mobility market [12]. The variation in the driver quantity for a long time, for example, is not considered. The total driver quantity in the ride-hailing market is assumed to be stable in this study. This can be further explored on the impact of the variation in the total driver quantity on the optimal commission rate. Second, the cooperation between RHPs is not considered. In practice, there usually exists not only competition but also cooperation between platforms, which is taken as a coopetition relationship in aggregated platform contexts [13].

Zhang et al. [14] addressed electric ride-hailing vehicles (ERVs), as a novel mobility paradigm, utilize intelligent platform-based scheduling to improve operational efficiency and create numerous employment opportunities. Contrasted with the traditional fuel-powered ride-hailing vehicle with the average annual carbon emissions of 4.6 tons, pure electric vehicles (PEVs) reduce emissions by 85.71 kg monthly during operation. Xu et al. [15] empirically studied drivers' response behaviour to ride-hailing requests based on data from ride-hailing platforms, and clarified that drivers are more likely to respond to requests with economic incentives. Utilizing ride-hailing and taxi order data from Xiamen.

3. Proposed Methodology

The proposed methodology presents a structured and data-driven framework for analysing and predicting driver attrition using a hybrid deep learning and machine learning approach. The system aims to identify drivers who are at high risk of leaving the organization, enabling proactive decision-making and effective retention strategies. The framework follows a comprehensive analytical pipeline that begins with data ingestion and progresses through preprocessing, feature transformation, model training, and prediction stages. The overall architecture of the system is illustrated in Figure. 2. By integrating advanced algorithms with statistical analysis, the system effectively captures complex relationships and hidden patterns within driver-related data. The framework is designed to handle structured, high-dimensional datasets and ensures reliable performance through comparative model evaluation. The system supports both single-instance prediction and batch-level analysis, making it suitable for real-time decision-making as well as large-scale organizational assessment.

User Interface (Input / Prediction Interface)

- Users interact with the system through a web-based interface built using the Flask framework, designed for both technical and non-technical stakeholders.
- The interface facilitates the input of driver-specific data for real-time single predictions or large-scale CSV-based batch uploads.
- Provides a centralized dashboard where users can view real-time prediction results, historical analytical insights, and live model performance metrics.

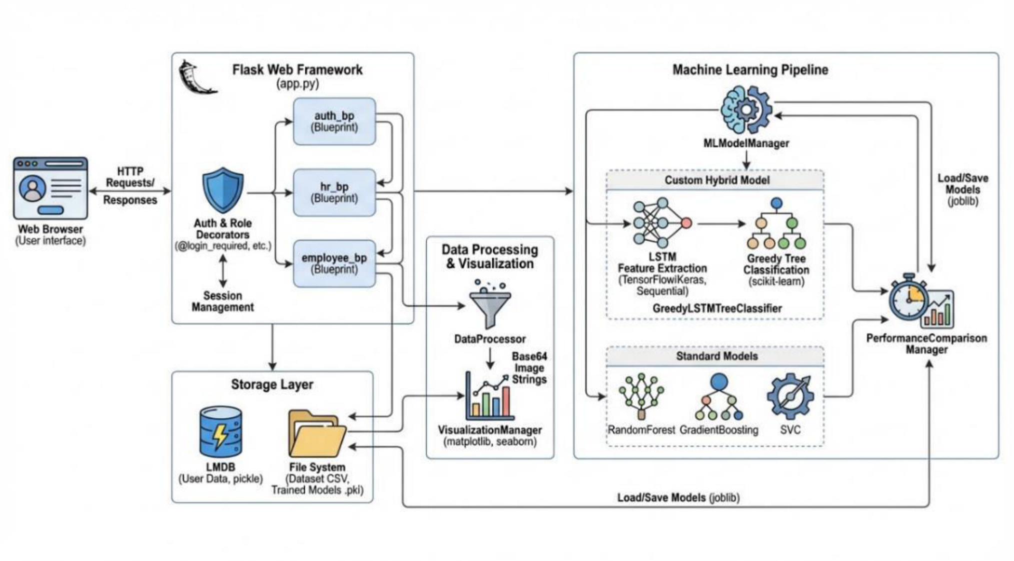


Figure. 2: Proposed system architecture.

Backend Processing Environment

- Acts as the orchestration layer, managing the secure flow of data across the preprocessing, training, and inference stages.
- Handles dynamic model loading and execution, ensuring high-speed result generation for both individual and batch queries.
- Maintains the integrity of the system by ensuring a seamless handshake between the raw input modules and the final analytical outputs.

Driver Dataset (Structured Data Source)

- Serves as the foundational data layer, containing structured attributes such as age, job role, income levels, and subjective satisfaction scores.
- Encapsulates a variety of categorical and numerical features that represent the complex environmental and personal factors influencing attrition.
- Provides the high-fidelity input required for deep feature extraction and robust classification.

Data Preprocessing Module

- Performs critical data cleansing by removing irrelevant noise and handling missing or inconsistent entries to prevent algorithmic bias.
- Applies Label Encoding to transform categorical variables into a mathematical format suitable for deep learning.
- Implements Standardization methods for feature scaling, ensuring that variables with different units contribute equally to the model's learning process.

Feature Extraction and Transformation

- Utilizes LSTM layers to extract deep latent features from the structured driver data, capturing non-linear relationships.
- Identifies hidden dependencies among work conditions and satisfaction levels that are often missed by traditional feature engineering.

- Produces an enhanced, high-dimensional representation of the data to boost the predictive accuracy of the subsequent classifiers.

Class Balancing and Data Preparation

- Addresses potential data skewness by ensuring a balanced representation of "Attrition" vs. "Non-Attrition" classes.
- Employs strategic data splitting into training and testing sets to ensure the model generalizes well to unseen driver profiles.
- Prepares the final vectorized datasets for unbiased, robust model learning and validation.

Classification Models and GLSTMT

- Implements a suite of baseline models, including RF, GB, and SVC for benchmarking.
- Introduces the proposed GLSTMT, which utilizes the LSTM-extracted features within a greedy tree-based learning structure.
- Conducts an independent execution of each algorithm to facilitate a rigorous comparative performance analysis.

Prediction Output and Interpretation

- Generates binary attrition predictions (Yes/No) alongside granular satisfaction-level outputs for each driver profile.
- Maps complex model probabilities into user-friendly formats, making the results interpretable for HR and management decision-making.
- Supports dual-mode reporting for both immediate individual assessments and comprehensive batch-level reports.

Visualization, Analytics, and Evaluation

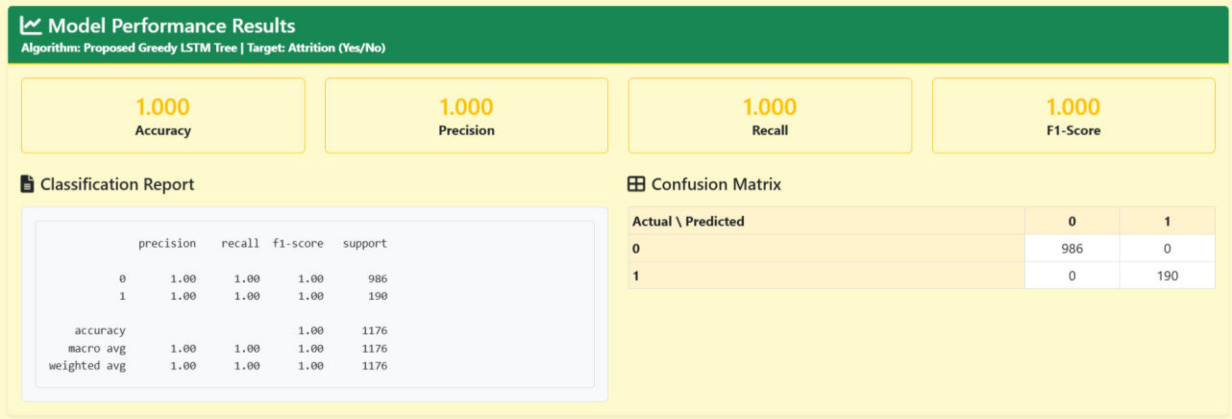
- Computes industry-standard metrics including Accuracy, Precision, Recall, and F1-score to validate model reliability.
- Provides a Visualization Module that renders correlation heatmaps, attrition distribution charts, and confusion matrices.
- Enhances organizational decision-making by revealing the underlying trends and patterns that drive workforce instability.

4. Result Description

The results of the research demonstrate the successful implementation of an AI-driven HR attrition prediction system integrated with a Flask-based web interface. The research provides a seamless workflow from user registration to HR analytics and real-time attrition prediction. Each component, including data visualization, model training, and predictions, is presented through interactive screens and dashboards. The system allows HR managers to analyse employee trends, understand key attrition factors, and make informed decisions based on predictive analytics. Performance evaluation of different models validates the effectiveness of the chosen algorithms in predicting employee turnover accurately.

Figure 3 (a) The Attrition classification report illustrates the model’s capability to accurately distinguish between drivers who remain in the organization and those who leave. The evaluation metrics such as precision, recall, and F1-score demonstrate the strong classification performance of the proposed model in attrition prediction.

Figure 3 (b) The Environment Satisfaction analysis evaluates the effectiveness of the proposed model in classifying satisfaction levels associated with the driver work environment. The classification report indicates accurate identification of satisfaction categories ranging from low to very high.



Model Performance Results
Algorithm: Proposed Greedy LSTM Tree | Target: Attrition (Yes/No)

1.000	1.000	1.000	1.000
Accuracy	Precision	Recall	F1-Score

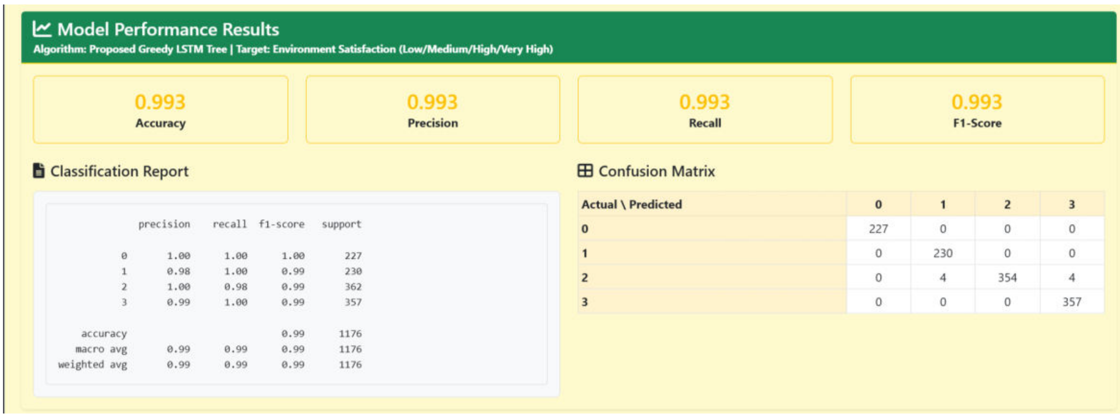
Classification Report

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       986
           1       1.00      1.00      1.00       190

    accuracy                           1.00      1176
   macro avg       1.00      1.00      1.00      1176
weighted avg       1.00      1.00      1.00      1176
```

Confusion Matrix

Actual \ Predicted	0	1
0	986	0
1	0	190

(a)



Model Performance Results
Algorithm: Proposed Greedy LSTM Tree | Target: Environment Satisfaction (Low/Medium/High/Very High)

0.993	0.993	0.993	0.993
Accuracy	Precision	Recall	F1-Score

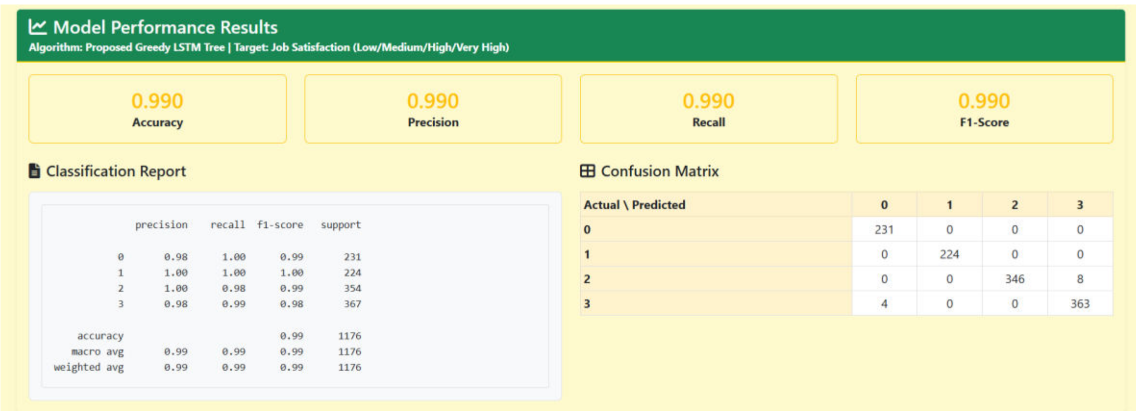
Classification Report

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       227
           1       0.98      1.00      0.99       230
           2       1.00      0.98      0.99       362
           3       0.99      1.00      0.99       357

    accuracy                           0.99      1176
   macro avg       0.99      0.99      0.99      1176
weighted avg       0.99      0.99      0.99      1176
```

Confusion Matrix

Actual \ Predicted	0	1	2	3
0	227	0	0	0
1	0	230	0	0
2	0	4	354	4
3	0	0	0	357

(b)



Model Performance Results
Algorithm: Proposed Greedy LSTM Tree | Target: Job Satisfaction (Low/Medium/High/Very High)

0.990	0.990	0.990	0.990
Accuracy	Precision	Recall	F1-Score

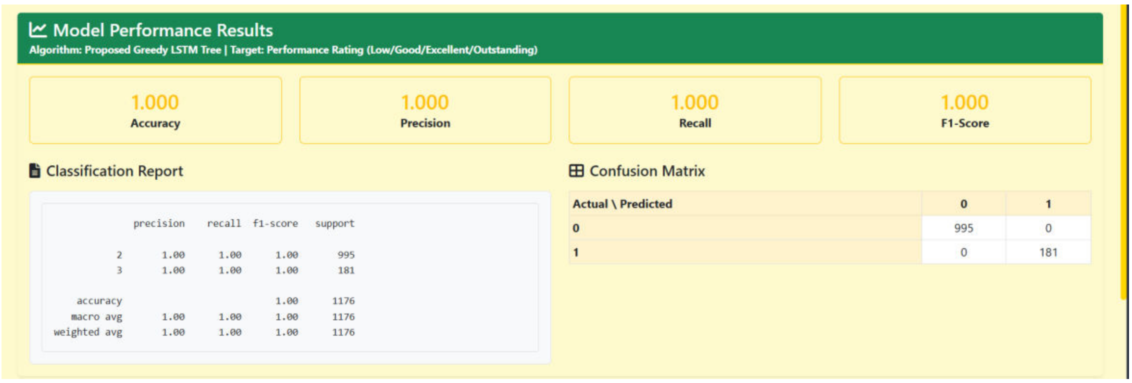
Classification Report

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       231
           1       1.00      1.00      1.00       224
           2       1.00      0.98      0.99       354
           3       0.98      0.99      0.98       367

    accuracy                           0.99      1176
   macro avg       0.99      0.99      0.99      1176
weighted avg       0.99      0.99      0.99      1176
```

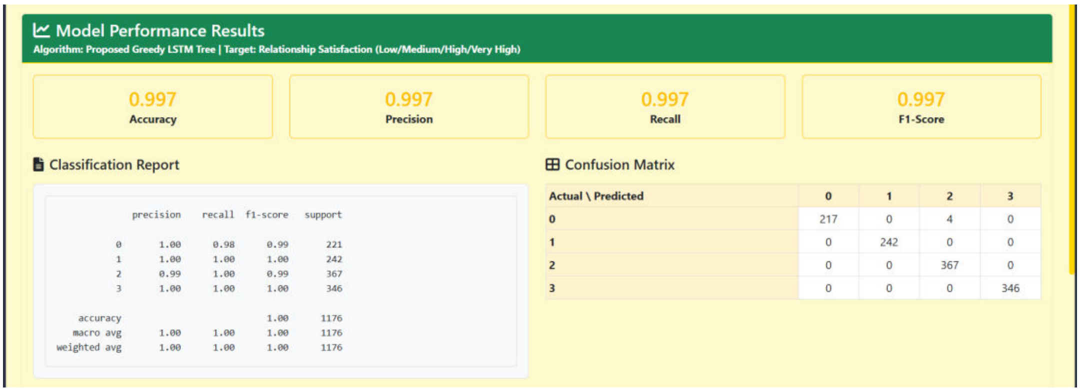
Confusion Matrix

Actual \ Predicted	0	1	2	3
0	231	0	0	0
1	0	224	0	0
2	0	0	346	8
3	4	0	0	363

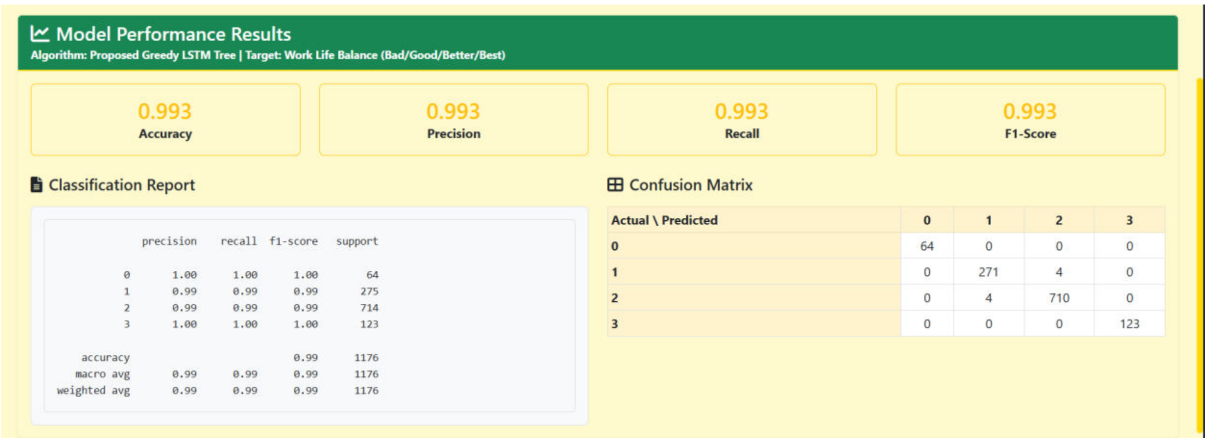
(c)



(d)



(e)



(f)

Figure. 3: Performance and classification reports of various target attributes from GLSTMT Model (a) Attrition, (b) Environment Satisfaction, (c) Job Satisfaction, (d) Performance Rating, (e) Relationship Satisfaction, (f) Work Life Balance.

Figure 3 (c) The Job Satisfaction report demonstrates the model's capability to classify driver satisfaction levels related to job roles, responsibilities, and engagement within the organization. The evaluation metrics reflect the strong predictive performance of the proposed hybrid model.

Figure 3 (d) The Performance Rating report evaluates the classification accuracy of the GLSTMT model in predicting driver performance categories. The results demonstrate precise identification of performance levels based on historical driver attributes.

Figure 3 (e) The Relationship Satisfaction report reflects the model's ability to classify variations in satisfaction related to relationships between drivers and management. The classification metrics highlight accurate identification of interpersonal satisfaction levels.

Figure 3 (f) The Work Life Balance report measures the model's capability to classify drivers according to their balance between professional responsibilities and personal life. The results demonstrate the strong performance of the proposed model in identifying work-life balance levels within the driver dataset.

Attrition

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Greedy LSTM Tree	1.0	1.0	1.0	1.0
RandomForestClassifier	0.8384	0.703	0.8384	0.7648
GradientBoostingClassifier	0.8384	0.703	0.8384	0.7648
SVC	0.8384	0.703	0.8384	0.7648

(a)

Environment Satisfaction

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Greedy LSTM Tree	0.9932	0.9933	0.9932	0.9932
RandomForestClassifier	0.3265	0.2973	0.3265	0.2483
GradientBoostingClassifier	0.3333	0.2167	0.3333	0.2221
SVC	0.3078	0.0948	0.3078	0.1449

(b)

Job Satisfaction

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Greedy LSTM Tree	0.9898	0.9899	0.9898	0.9898
RandomForestClassifier	0.3367	0.2095	0.3367	0.24
GradientBoostingClassifier	0.3163	0.2028	0.3163	0.1885
SVC	0.3121	0.0974	0.3121	0.1485

(c)

Performance Rating

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Greedy LSTM Tree	1.0	1.0	1.0	1.0
RandomForestClassifier	0.8461	0.7159	0.8461	0.7755
GradientBoostingClassifier	0.8461	0.7159	0.8461	0.7755
SVC	0.8461	0.7159	0.8461	0.7755

(d)

Relationship Satisfaction

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Greedy LSTM Tree	0.9966	0.9966	0.9966	0.9966
RandomForestClassifier	0.3401	0.4213	0.3401	0.2314
GradientBoostingClassifier	0.3206	0.2592	0.3206	0.169
SVC	0.3121	0.0974	0.3121	0.1485

(e)

Work Life Balance

Algorithm	Accuracy	Precision	Recall	F1 Score
Proposed Greedy LSTM Tree	0.9932	0.9932	0.9932	0.9932
RandomForestClassifier	0.6071	0.3686	0.6071	0.4587
GradientBoostingClassifier	0.6071	0.3686	0.6071	0.4587
SVC	0.6071	0.3686	0.6071	0.4587

(f)

Figure. 4: Model comparison screen. (a) Attrition, (b) Environment Satisfaction, (c) Job Satisfaction, (d) Performance Rating, (e) Relationship Satisfaction, (f) Work Life Balance.

Figure 4 (a) The Attrition comparison screen shows the performance results of all models for predicting whether drivers remain in the organization or leave. The comparison highlights differences in accuracy, precision, recall, and F1-score values among the models, demonstrating the effectiveness of the proposed GLSTMT model in identifying driver attrition patterns.

Figure 4 (b) The Environment Satisfaction comparison screen evaluates the classification performance of the models in identifying driver satisfaction levels related to the working environment. The comparison illustrates how each algorithm performs in distinguishing between different satisfaction categories ranging from low to very high.

Figure 4 (c) The Job Satisfaction comparison screen displays the performance of the machine learning models in predicting job satisfaction levels among drivers. The results provide a clear view of how effectively each model classifies satisfaction levels associated with job roles, work conditions, and organizational engagement.

Figure 4 (d) The Performance Rating comparison screen illustrates the classification results of different models in predicting driver performance levels. The comparison highlights the ability of each model to accurately identify performance categories based on driver performance attributes.

Figure 4 (e) The Relationship Satisfaction comparison screen evaluates the models based on their performance in identifying the level of satisfaction related to relationships between drivers and organizational management. The displayed metrics provide insights into the classification capability of each model for interpersonal satisfaction factors.

Figure 4 (f) The Work Life Balance comparison screen presents the comparative performance of the models in classifying drivers according to their balance between professional responsibilities and personal life. The results demonstrate the effectiveness of the proposed model in identifying different work-life balance levels within the dataset.

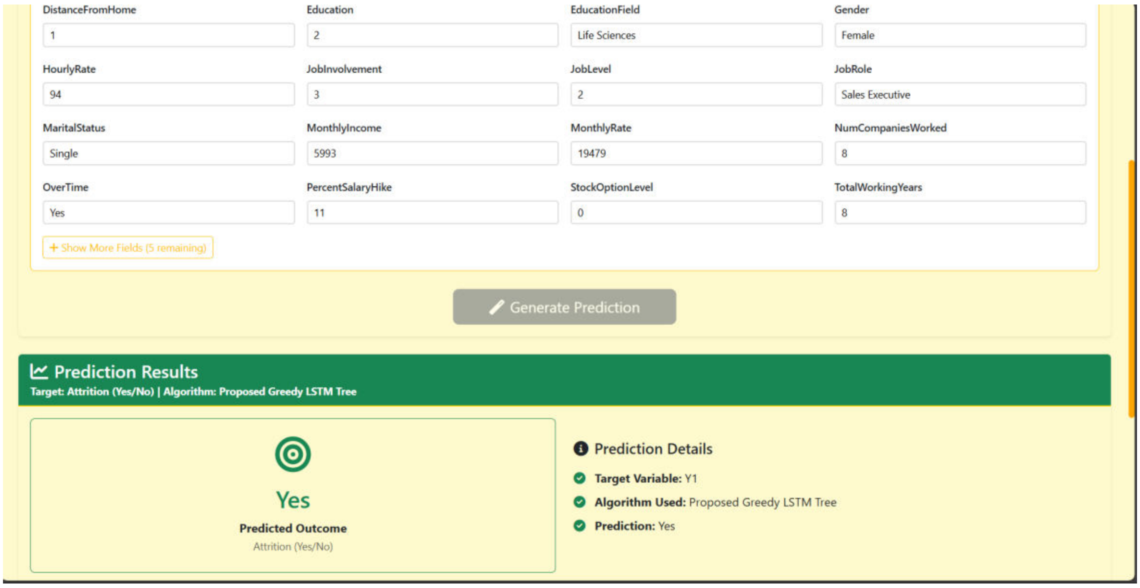


Figure. 5: Predictions on test data.

The figure 5 illustrates the prediction outcome of an employee attrition analysis system, highlighting the final classification generated by the proposed GLSTMT algorithm. It depicts the successful processing of multiple input features related to employee demographics, job role, and work conditions to determine attrition status. The visualization presents a clear indication of the predicted result as “Yes,” signifying the likelihood of employee attrition based on the given attributes. It also showcases the integration of target variable identification and model inference within the prediction pipeline. The figure emphasizes the interpretability of the model by displaying concise prediction details, including the algorithm used and the evaluated output. Furthermore, it reflects the effectiveness of the hybrid learning approach in capturing complex patterns from structured HR data for decision support.

5. CONCLUSION

This research successfully designed and implemented an AI-driven Employee Attrition Prediction System using Machine Learning and Deep Learning techniques. The system integrated data preprocessing, exploratory data analysis, and multiple predictive models including SVC, RF, GB, and GLSTMT. A Flask-based web application provided secure user authentication, HR analytics dashboards, and real-time prediction functionality. The experimental results demonstrated that ensemble and hybrid models achieved higher prediction accuracy compared to traditional classifiers. Visual analytics through EDA improved understanding of key attrition factors such as job satisfaction, work-life balance, and years at the company. The system enabled HR professionals to identify high-risk employees early and supported data-driven workforce management decisions. The research delivered an efficient, scalable, and intelligent solution for employee attrition analysis in real-world organizational environments.

REFERENCES

[1] Chen, X.; Deng, H. Improving Ride-Hailing Platform Operations in Dynamic Markets: A Drivers’ Switching Perspective. Systems 2025, 13, 80. https://doi.org/10.3390/systems13020080

[2] Cohen, M.C.; Zhang, R. Competition, and coopetition for two-sided platforms. Prod. Oper. Manag. 2020, 31, 1997–2014.

[3] Sun, S.; Ertz, M. Dynamic evolution of ride-hailing platforms from a systemic perspective: Forecasting financial sustainability. Transp. Res. Part C Emerg. Technol. 2021, 125, 103003.

[4] Zhong, Y.; Yang, T.; Cao, B.; Cheng, T.C.E. On-demand ride-hailing platforms in competition with the taxi industry: Pricing strategies and government supervision. Int. J. Prod. Econ. 2022, 243, 108301.

[5] Sun, L.; Teunter, R.H.; Hua, G.; Wu, T. Taxi-hailing platforms: Inform or Assign drivers? Transp. Res. Part B Methodol. 2020, 142, 197–212.

[6] Xu, Z.; Yin, Y.; Chao, X.; Zhu, H.; Ye, J. A generalized fluid model of ride-hailing systems. Transp. Res. Part B Methodol. 2021, 150, 587–605.

[7] Ke, J.; Qian, S. Leveraging ride-hailing services for social good: Fleet optimal routing and system optimal pricing. Transp. Res. Part C Emerg. Technol. 2023, 155, 104284.

[8] Bandiera, C.; Connors, R.D.; Viti, F. Mobility service providers' interacting strategies under multi-modal equilibrium. Transp. Res. Part C Emerg. Technol. 2024, 168, 104766.

[9] Xi, H.; Li, M.; Hensher, D.A.; Xie, C.; Gu, Z.; Zheng, Y. Strategizing sustainability and profitability in electric Mobility-as-a-Service (E-MaaS) ecosystems with carbon incentives: A multi-leader multi-follower game. Transp. Res. Part C Emerg. Technol. 2024, 166, 104758.

[10] Yao, R.; Zhang, K. Design an intermediary mobility-as-a-service (MaaS) platform using many-to-many stable matching framework. Transp. Res. Part B Methodol. 2024, 189, 102991.

[11] Sun, S.; Wang, Z.; Wang, W. The impact of regulatory policy on the growth of ride-hailing platform: System dynamics perspective. Phys. A Stat. Mech. Appl. 2023, 632, 1–21.

[12] Chen, X.G.; Deng, H.; Guan, S.Q.; Han, F.X.; Zhu, Z.H. Cooperation-oriented multi-modal shared mobility for sustainable transport: Developments and challenges. Sustainability 2024, 16, 11207.

[13] Ma, L.; Tao, Z.; Wei, Q.; Huo, B. Cooperate with aggregation platform or not? Optimal decision for the on-demand ride service platform. Res. Transp. Bus. Manag. 2025, 59, 101251.

[14] Lyu, W.; Hu, Y.; Liu, J.; Chen, K.; Liu, P.; Deng, J.; Zhang, S. Impact of battery electric vehicle usage on air quality in three Chinese first-tier cities. Sci. Rep. 2024, 14, 21.

[15] Xu, K.; Sun, L.; Liu, J.; Wang, H. An empirical investigation of taxi driver response behavior to ride-hailing requests: A spatio-temporal perspective. PLoS ONE 2018, 13, e0198605.