IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 February 2024, accepted 16 March 2024, date of publication 22 March 2024, date of current version 6 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380895

RESEARCH ARTICLE

An Automated Approach for Predicting Road Traffic Accident Severity Using Transformer Learning and Explainable AI Technique

OMAR IBRAHIM ABOULOLA[1], EBTISAM ABDULLAH ALABDULQADER[2], AISHA AHMED ALARFAJ[3], SHTWAI ALSUBAI[4], AND TAI-HOON KIM[5], (Member, IEEE)

[1]College of Computer Science and Engineering, University of Jeddah, Jeddah 21589, Saudi Arabia
[2]Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh 11421, Saudi Arabia
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[5]School of Electrical and Computer Engineering, Chonnam National University, Yeosu-si, Jeollanam-do 59626, Republic of Korea

Corresponding authors: Aisha Ahmed Alarfaj (aiaalarfaj@pnu.edu.sa) and Tai-Hoon Kim (taihoonn@chonnam.ac.kr)

This work was supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, through Princess Nourah bint Abdulrahman University Researchers Supporting Project under Grant PNURSP2024R348.

ABSTRACT Traffic accidents continue to be a significant cause of fatalities, injuries, and considerable disruptions on our highways. Understanding the underlying factors behind these incidents is crucial for improving safety on road networks. While recent studies have highlighted the usefulness of predictive modeling in uncovering factors leading to accidents, there remains a gap in explaining the inner workings of complex machine learning and deep learning models and how various features influence accident prediction. This lack of transparency may lead to these models being perceived as black boxes, potentially undermining trust in their findings among stakeholders. The primary aim of this research is to develop predictive models using diverse transfer learning techniques and shed light on the most influential factors using Shapley values. In predicting injury severity in accidents, we employ Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Residual Networks (ResNET), EfficientNetB4, InceptionV3, Extreme Inception (Xception), Visual Geometry Group (VGG19), AlexNet, and MobileNet. Among these models, MobileNet emerges with the highest accuracy at 0.9817. Furthermore, by comprehending how different features impact accident prediction models, researchers can deepen their understanding of the factors contributing to accidents and devise more effective interventions for their prevention.

INDEX TERMS Intelligent transportation system, road accidents severity, MobileNet, explainable AI (XAI).

I. INTRODUCTION

The escalating number of automobiles on roads has become a pressing issue in contemporary society, coinciding with a parallel increase in traffic accidents, resulting in substantial human and economic tolls [1]. Annually, a significant number of individuals suffer injuries or fatalities in vehicular accidents worldwide, resulting in substantial human and economic setbacks. To efficiently alleviate the harm and losses stemming from road traffic accidents (RTAs), understanding the underlying factors of these events and the extent of resulting injuries is crucial. Considering the increasing intricacy of road systems and the growing number of vehicles, it is vital to embrace a data-centric strategy for examining patterns in accidents and pinpointing risk elements. Continual investigation and comprehension of RTA causes, coupled with the implementation of effective preventive measures, are paramount [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Kah Phooi (Jasmine) Seng.

A recent study conducted by the World Health Organization (WHO) emphasizes the worldwide importance of RTAs, contributing to more than 1.3 million fatalities each year. Globally, vehicular accidents stand as the leading cause of mortality among adolescents and young adults [3]. The intensity of road incidents acts as a pivotal gauge for injuries associated with accidents, with diverse elements contributing to incidents of varying seriousness [4]. Despite advancements, there has been no substantial reduction in traffic accident fatalities and injuries over the past two decades. Predictive models offer a proactive approach to addressing accident factors, potentially curbing fatalities, reducing costs, and enhancing comprehension. Weather conditions have been discussed in relation to various road types [5], alongside other significant factors such as lighting conditions, road class, vehicle count, and traffic volume [6].

The primary objective of accident data analysis is to identify key factors influencing the occurrence of RTAs, thereby addressing critical road safety concerns. The efficacy of accident prevention strategies hinges on the authenticity of collected and estimated data, as well as the appropriateness of chosen analytical methodologies [7]. Selecting suitable data analysis methods is vital for discerning accident causes in specific zones or study locations and accurately predicting daily accident probabilities or assessing safety levels for different road user groups in those areas [8]. Consequently, research quality is contingent upon methodological selection. Machine learning approaches have been utilized to predict traffic accidents [9], with Zhang et al. [10] utilizing generalized random forests to estimate heterogeneous treatment effects in road safety studies, providing comprehensive information to local authorities and policymakers to enhance speed camera program effectiveness. Some researchers have employed statistical methods [11], reinforcement learning approaches [12], hybrid models [13], and deep learning models [14]. Zhao et al. [15] employed a deep convolutional neural network and random forest for accident risk prediction.

A. RESEARCH GAP AND MOTIVATION

Despite numerous attempts to investigate accident-contributing factors, limited attention has been given to explaining black box models [16]. The authors explored five machine learning models alongside explainable machine learning [17]. This study aims to develop an accident injury severity prediction model based on transfer learning and identify major contributing factors using an explainable approach. The US accident dataset (2016-2021) is employed for predicting traffic accident severity, with the objective of creating an automated system for categorizing accident severity. The motivation behind the research on Traffic Accident Severity Prediction Using MobileNet Transfer Learning Model and SHAP XAI Technique stems from the pressing need to enhance road safety and mitigate the adverse impacts of traffic accidents. With traffic accidents remaining a significant cause of fatalities, injuries, and economic losses worldwide, there is a critical demand for effective predictive models to aid in accident severity prediction and prevention efforts. Traditional approaches often lack the predictive power and interpretability required for accurate and actionable insights. Hence, the motivation for this research lies in harnessing the capabilities of advanced machine learning techniques, such as transfer learning with MobileNet architecture, coupled with Explainable AI (XAI) techniques like SHAP. By leveraging state-of-the-art methodologies, the research aims to develop more accurate, interpretable, and actionable models for traffic accident severity prediction. Ultimately, the overarching goal is to empower stakeholders, including transportation planners, law enforcement agencies, and policymakers, with the tools and insights needed to implement targeted interventions and enhance road safety outcomes.

B. NOVELTY OF THE PROPOSED WORK

The study introduces novel elements to traffic accident severity prediction, contributing to road safety. Innovations include adopting MobileNet for transfer learning, providing accuracy and efficiency. A comprehensive model comparison explores diverse models, enhancing understanding and highlighting transfer learning adaptability. The inclusion of the SHapley Additive exPlanations (SHAP) model addresses model interpretability, offering transparency in predicting severity. Rigorous generalizability testing confirms the model's robustness beyond the initial dataset, emphasizing practical applicability. The research advances predictive modeling in road safety, offering an accurate, interpretable, and widely applicable solution for mitigating traffic accident impacts. This study makes the following noteworthy contributions:

- Utilization of a MobileNet model based on transfer learning, exhibiting exceptional accuracy in road traffic accident severity prediction.
- Conducting experiments on three deep learning models (Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM)) and five transfer learning models (ResNET, EfficientNetB4, InceptionV3, Xception, and MobileNet).
- Demonstrating the significance of various features through the utilization of the SHapley Additive exPlanations (SHAP) model.
- Testing the proposed model on another dataset to validate its generalizability.

The organization of this research is outlined as follows: Section II offers a glimpse into previous studies in this domain. Section III unveils the suggested methodology and delineates the deep learning and transfer learning models. Section IV lays out the evaluation of the proposed approach, encompassing experimental outcomes and pertinent discussions. Ultimately, Section V functions as the culmination of this study.

II. RELATED WORK

In the past few years, the utilization of machine learning has surged in predicting the severity of accidents due to its capacity to reveal concealed associations and offer more precise insights compared to conventional statistical approaches. These statistical approaches for forecasting accident severity face limitations like diminished accuracy and impractical estimation. Researchers have turned to machine learning and deep learning methodologies to enhance prediction efficacy. This section presents an overview of previous methodologies employed for forecasting traffic accident severity.

Regarding traffic accident characteristics, Gan et al. [18] utilized a random forest approach to identify eight attributes in traffic accident data for predicting accident severity. Factors such as engine capacity, time of day, vehicle age, month, day of week, driver age group, vehicle movement, and speed restrictions were considered. In another research [19], several machine learning models including Naive Bayes (NB), Random Forest (RF), Adaptive Boosting (ADA), and Logistic Regression (LR) were evaluated for forecasting the severity of injuries in RTAs. The highest results have been achieved by RF with 75% accuracy.

In Saudi Arabia, Aldhari et al. [20] suggested a machine learning approach to predict road accident severity, employing three models: RF, LR, and XGBoost, with SHAP used to address bias concerns. XGBoost achieved the 71% accuracy score for binary classification and 94% for multi-class classification. The authors [21] devised a method to forecast the seriousness of accidents employing deep-learning methodologies such as Multilayer Perceptron (MLP), and Recurrent Neural Network (RNN). The highest results have been achieved by the RNN model with 71.7% of accuracy.

Wahab and Jiang [22] proposed a basic CART model for predicting the seriousness of motorcycle accidents, along with the use of PART and MLP models. The CART attained a 73.8% accuracy, while the PART model scored 73.45%. The authors [23] suggested a deep neural network for predicting RTAs for the Internet of Vehicles, employing various models. The DNN surpassed alternative models and demonstrated proficiency in both the initial and subsequent phases of clustering.

Jamal et al. [24] introduced a network utilizing various machine learning models to improve the result in forecasting the severity of road accidents. The XGBoost model outperformed others in classwise accuracy and overall prediction performance, achieving an outstanding 93% accuracy. Manzoor et al. [25] proposed RFCNN, an ensemble learning model combining machine and deep learning models to identify relevant reasons for RTA's severity. The proposed RFCNN model demonstrated good accuracy on the 20 most relevant characteristics. Bahiru et al. [26] evaluated multiple machine learning methods including ID3, CART, J48, and NB with 96% accuracy by J48 model. Cicek et al. [17] employed various machine learning models with explanations to predict accident severity, while Yang et al. [16] applied deep learning for multitasking and predicted severity levels of traffic accidents with explanations. They conducted experiments on a Chinese traffic accident dataset. Although many researchers have utilized machine learning and deep learning for predicting traffic accident severity, few have conducted comparative analyses of different deep learning methods. Additionally, limited research has explored contributing factors using explanations. The incorporation of explanation techniques enhances transparency, interpretability, explanatory capacity, domain knowledge integration, and scientific coherence of models [27], addressing the common perception of prediction methods as black boxes. This research work makes use of the SHAP XAI technique to explain to what extent which feature contributes to making specific target class predictions. Therefore, this study selects five distinct transfer learning methods and compares their predictive capabilities. As a novel contribution, an explainable technique is applied to forecast the most influential factors contributing to accidents in the proposed models. A summary of prior studies is presented in Table 1.

III. DATASET AND METHODOLOGY

This section provides a comprehensive overview of the dataset utilized, the deep learning models, transfer learning models employed, and the parameters utilized for evaluating the performance of these models in predicting the severity of traffic accidents. Additionally, the experimental framework adopted is depicted in Figure 1.

A. DATASET

This study utilizes accident data spanning a five-year period (2016-2020) from New Zealand, sourced from the Crash Analysis System (CAS) maintained by the Te Manatū Waka Ministry of Transport. The dataset is publicly available through the open data portal. Two distinct sets of data were obtained from the CAS system, containing details related to individuals involved, vehicles, and accident particulars. These datasets, referred to as the 'person' dataset and the 'accident' dataset, were amalgamated to form a comprehensive dataset focusing on factors contributing to accidents. Initially, the combined dataset comprised 378,820 entries across 101 columns. However, several columns, deemed irrelevant to accident causation, were excluded from the study. For instance, a column containing information regarding neighboring police stations was deemed superfluous. As a result, 36 features relevant to various aspects of accidents were selected for analysis, including crash type, crash location characteristics, environment, vehicle types, and personal factors influencing accident severity. Accident types are categorized based on severity into four levels, as outlined in Table 2.

B. MULTILAYER PERCEPTRON (MLP)

The multilayer perceptron model [28] represents a significant advancement over the original perceptron model devised by Rosenblatt. Unlike the perceptron, which was constrained

TABLE 1. Overview of prior studies on predicting accident severity.

Ref	Methods	Key Findings	Limitations
[18]	RF, Light-GBM	Identified 8 attributes for accident validation.	Data limitations, small dataset
[19]	NB, RF, ADA, LR	RF model achieved highest accuracy rate of 75.5%.	Limited comparison
[20]	RF, LR, XGBoost, SHAP	Machine learning approach in Saudi Arabia, XGBoost achieved highest accuracy.	Limited discussion of results
[21]	MLP, BLR, RNN	RNN model achieved accuracy of 71.77%.	Not handling imbalanced dataset
[22]	CART, PART, MLP	Basic CART model for motorcycle accident severity prediction achieved 73.45% accuracy (PART).	Lack of feature importance analysis
[23]	DNN, DT C4.5, NB, DBN, MLP, Bayesian Network	DNN outperformed other models in predicting accident risk.	Limited explanation of model selection
[24]	RF, LR, DE, XGBoost	XGBoost achieved highest accuracy of 93% and identified influential elements.	Limited model comparison, no feature importance analysis
[25]	RFCNN	Ensemble learning model achieved high accuracy.	Feature selection may lead to overfitting
[26]	ID3, NB, J48, CART	J48 machine learning model achieved 96% accuracy.	Limited model comparison
[16]	DNN	Proposed model predicts accident severity with explanation.	Limited models used for comparison
[17]	DT, NB, MLP, SVM, NN	ANN-MLP achieved 76.90% accuracy.	Low accuracy results

TABLE 2. Description of the dataset.

Sr. No.	Severity Class	Description	Number of Records
1	Fatal Crash	A road accident resulting in loss of life.	1543
2	Serious Injury Crash	A road accident where one or more parties required medical attention and were transported to a hospital.	10582
3	Minor Injury Crash	A road accident where no one required medical attention but sustained minor injuries like bruises and superficial cuts.	42888
4	Non-Injury Crash	A road accident where no injuries were sustained, and the presence of law enforcement may not always be required.	129304

to handling linearly separable problems in basic logic, the multilayer perceptron introduces multiple layers of functional neurons, enabling it to address nonlinear separable problems. Its architecture comprises fully interconnected layers, facilitating the organized flow of information. The training utilizes the error back-propagation algorithm to minimize the cumulative error on the training set, typically measured using mean-square error (MSE) for each sample.

C. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN [29] is a deep neural network specifically engineered for tasks such as image recognition, classification, and segmentation. It incorporates convolution, non-linear activation, and pooling layers to extract features. Stacked CNNs are utilized for specialized tasks, such as detecting parasites in infected cell images. The architecture is multi-layered, with each layer applying filters or kernels to input data to generate feature maps. Convolutional layers' outputs are concatenated and passed into fully connected layers for further analysis. CNN has become a standard in medical domain classification and employs secure interaction protocols for privacy-preserving feature extraction.

D. LONG SHORT-TERM MEMORY (LSTM)

Long Short-Term Memory (LSTM) [30] represents a specialized recurrent neural network (RNN) architecture developed to overcome traditional RNNs' limitations in handling long-term dependencies within sequential data. LSTMs excel in tasks involving sequences, including natural language processing, time series analysis, and speech recognition. Key features include mitigating the vanishing gradient problem in standard RNNs, memory cells for information storage and erasure, and gating mechanisms (input, forget, and output gates) to regulate data flow. LSTMs employ activation functions to analyze incoming data and train using "Backpropagation through Time" (BPTT). Variants like Bidirectional LSTMs and simpler Gated Recurrent Unit (GRU) networks have also been introduced. LSTMs find applications in various domains, from language modeling and stock price prediction to speech recognition and image captioning.

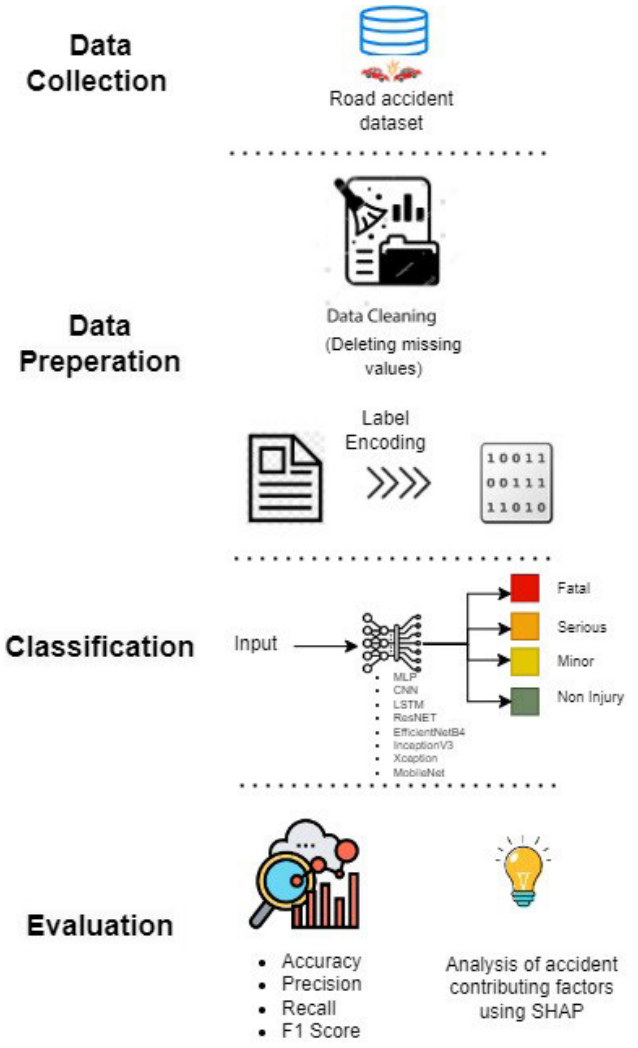

FIGURE 1. Architecture of the proposed framework.

E. RESNET

Residual Networks or ResNets [31], represent a robust and innovative type of deep neural network design that has had a significant impact on computer vision and deep learning since their inception in 2015 by Kaiming He et al. Developed to address the vanishing gradient problem hindering deep network training, ResNets' essential innovation lies in the ''residual block.'' This block incorporates two key routes: the identity path, reflecting the original input and transferring it directly to the output, and the residual path, applying a series of convolutional layers and non-linear activations to the input. Skip connections, also known as shortcut connections, facilitate gradient flow during training, enabling the training of extremely deep networks with hundreds or thousands of layers without performance degradation. ResNets have excelled in various image classification tasks, particularly in the ImageNet Large Scale Visual Recognition Challenge, and are widely employed in transfer learning. Over time, adaptations like ResNetV2 and Wide ResNets have significantly enhanced both performance and efficiency.

F. EFFICIENTNETB4

EfficientNetB4 [32] belongs to the EfficientNet family of neural networks, known for their exceptional performance in image classification tasks while remaining computationally efficient. Balancing model size, computational requirements, and accuracy, EfficientNetB4 employs a systematic approach to scale neural network architectures, achieving an optimal balance of depth, width, and resolution through compound scaling. It utilizes depth-wise separable convolutions and squeeze-and-excite blocks to improve efficiency and feature capture. Demonstrating top-tier accuracy on benchmarks like ImageNet while maintaining computational efficiency, EfficientNetB4 is widely used for transfer learning tasks.

G. INCEPTIONV3

InceptionV3 [33] has been widely utilized for image recognition tasks, achieving high accuracy with numerous convolutional, pooling, and activation layers. Its architecture incorporates inception modules, facilitating the network to acquire unique feature representations across various scales. Techniques, like factorized 1×1 convolutions and normalizing batches, are employed to reduce parameters and improve training process. While versatile across various tasks and datasets, InceptionV3 can be computationally intensive and memory-consuming.

H. XCEPTION

Xception [34], derived from ''extreme inception,'' is a deep CNN architecture proposed by François Chollet in 2017. It extends Inception's concepts by employing depthwise separable convolutions, which are more efficient. Consisting of depthwise and pointwise convolutions, depthwise separable convolutions reduce computational complexity. Known for its deep architecture capable of learning complex features, Xception excels in image classification accuracy.

I. ALEXNET

AlexNet, a pioneering convolutional neural network (CNN) design, initially gained acclaim for its remarkable achievements in image classification endeavors, notably excelling on the ImageNet dataset [35]. Nevertheless, its utility transcends conventional image classification realms, showcasing effectiveness in addressing challenges associated with categorical data classification. Harnessing its adeptness in hierarchical feature extraction, AlexNet demonstrates proficiency in processing categorical data inputs, capturing nuanced relationships embedded within the data. Through tailored adaptations of AlexNet's architecture and training methodologies, researchers have effectively applied it to a wide array of categorical data classification tasks, spanning text classification, sentiment analysis, and medical diagnosis.

J. VGG19

VGG19, an iteration of the VGG (Visual Geometry Group) framework, stands out for its intricate convolutional neural network structure comprising 19 layers [36]. Initially devised for image classification endeavors, VGG19 has transcended traditional computer vision realms, showcasing its prowess in navigating challenges associated with categorical data classification. Its deep-seated architecture facilitates the extraction of intricate hierarchical features from intricate categorical datasets, enabling precise differentiation among diverse classes. Leveraging transfer learning and fine-tuning methodologies, VGG19 can be tailored to suit various categorical data classification tasks, including those in natural language processing, sentiment analysis, and medical diagnosis, often delivering cutting-edge performance benchmarks.

K. MOBILENET

MobileNet [37] is designed for embedded devices with limited processing capabilities, efficiently balancing accuracy and model size. Its key innovation lies in the utilization of depthwise separable convolutions, dividing convolutions into depthwise and pointwise stages to significantly reduce computational costs and model size. MobileNet comes in multiple versions, including MobileNetV1, V2, and V3, each improving performance and efficiency. Widely used in mobile and embedded applications like object recognition and image classification, MobileNet showcases an innovative approach to creating efficient yet high-performing convolutional neural networks.

L. EVALUATION PARAMETERS

This study employs various evaluation criteria, including accuracy, F1 score, recall, and precision, to assess the effectiveness of transfer learning models. Additionally, confusion matrices are utilized to evaluate these algorithms' performance. A confusion matrix, also known as an error matrix, provides a tabular representation illustrating the classifier's performance on testing data, providing a graphical representation of the algorithm's efficacy.

Positive instances correctly predicted are denoted as True Positives (TP), whereas accurate predictions for the negative class are represented by True Negatives (TN). False Positives (FP) indicate erroneous predictions for the positive class when the actual class is negative, and False Negatives (FN) signify inaccurate predictions for the negative class when the true class is positive.

The model's comprehensive predictive precision is ascertained by appraising the ratio of accurate predictions to the entire instances in the dataset. Precision gauges the ratio of correctly predicted positive instances among all instances labeled as positive by the model. Recall alternatively termed the genuine positive rate or sensitivity, measures the model's efficacy in precisely capturing positive cases. The F1 measure, symbolizing the harmonic mean of precision and recall, provides an equitable evaluation of the model's comprehensive effectiveness.

TABLE 3. Results of Deep Learning Models for Traffic Accident Severity Detection.

Models	Accuracy	Precision	Recall	F1 score
MLP	0.8727	0.8229	0.8361	0.8258
CNN	0.8937	0.8663	0.8867	0.8719
LSTM	0.8127	0.8024	0.8329	0.8267

IV. RESULTS AND DISCUSSION

In this investigation, open-source libraries like Keras and TensorFlow were employed for the creation of pre-trained models. The Python language and the Anaconda platform were utilized for the examination of traffic accident severity through transfer learning algorithms. A Dell Poweredge T430 server, furnished with a GPU, managed the computational requirements of the dataset. This server featured eight cores, 32GB of RAM, and sixteen logical processors. The paper suggests employing transfer learning techniques to address the challenge of predicting traffic accidents, and various scientific approaches will be employed to assess the efficacy and significance of the proposed methodology.

A. RESULTS OF DEEP LEARNING MODELS FOR TRAFFIC SEVERITY PREDICTION

Table 3 presents a comparative evaluation of the performance of three deep learning models—MLP, CNN, and LSTM—in predicting traffic accident severity. Results indicate that CNN outperformed other deep learning models across evaluation measures, achieving 0.8937 accuracy, 0.8663 precision, 0.8867 recall, and 0.8719 F1 score. Following CNN, MLP achieved 0.8727 accuracy, 0.8229 precision, 0.8361 recall, and 0.8258 F1 score. LSTM performed least effectively among the models, with an accuracy score of 0.8127. The CNN model demonstrated superior performance in predicting traffic accident severity compared to other deep learning models.

B. RESULTS OF TRANSFER LEARNING MODELS FOR TRAFFIC SEVERITY PREDICTION

Table 4 provides a detailed analysis of various transfer learning models' performance in predicting traffic accident severity. It showcases the performance of different transfer learning models, including ResNET, EfficientNetB4, InceptionV3, Xception, and MobileNet. Results highlight MobileNet as the top performer, achieving the highest accuracy at 0.9817, alongside 0.9834 precision, 0.9891 recall, and 0.9848 F1 score. In contrast, Xception ranks lower in precision (0.8263) and F1 score (0.8519), suggesting areas for improvement. InceptionV3 and EfficientNetB4 achieved accuracy scores of 0.9248 and 0.9367, respectively. ResNET exhibited the second-highest results with 0.9527 accuracy, 0.9625 precision, 0.9819 recall, and 0.9767 F1 score.

These findings offer valuable insights for researchers and practitioners exploring transfer learning for traffic

TABLE 4. Outcomes from Transfer Learning Models in Detecting Traffic Accident Severity.

Models	Accuracy	Precision	Recall	F1 score
ResNET	0.9527	0.9625	0.9819	0.9767
EfficientNetB4	0.9367	0.8969	0.8832	0.8855
InceptionV3	0.9248	0.9387	0.9799	0.9698
Xception	0.9137	0.8263	0.8967	0.8519
AlexNet	0.9056	0.7923	0.8055	0.7989
VGG19	0.9459	0.8559	0.8847	0.8698
MobileNet	0.9817	0.9834	0.9891	0.9848

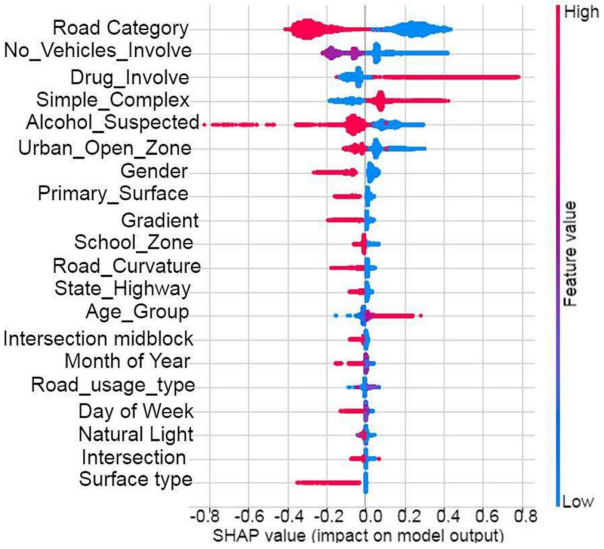

FIGURE 2. Impact of features on the performance of proposed model.

accident severity prediction, with MobileNet emerging as a particularly promising candidate for further real-world applications.

C. SHAP EXPLANATION

SHAP (SHapley Additive exPlanations) [38] is a widely adopted technique for elucidating the predictions made by machine learning models, encompassing even black-box models such as deep neural networks employed in transfer learning. The predominant variant of SHAP utilized for explicating black-box models is known as "Kernel SHAP." With its robust processing capabilities and visualization tools, researchers have increasingly leveraged SHAP for investigating road safety [39], [40].

In this investigation, Shapley values for each feature are computed using the Python Shap library to unveil the importance of features in the black-box transfer learning model. SHAP underscores the significance of features in predicting water quality. While SHAP's feature relevance surpasses that of traditional methods, its standalone usage only yields limited additional insights.

The SHAP plot organizes features in a descending order, showcasing their significance. Features with high importance are depicted in red at the top of the plot, while those with low importance are represented in blue at the bottom along the Y-axis. The X-axis illustrates the influence of these features on the model's output. Each point on the SHAP plot corresponds to a data point from the training dataset. When an X-axis value is positioned to the left of 0, it indicates an observation that shifts the target value in a negative direction, whereas a value to the right of 0 shifts it positively.

As illustrated in Figure 2, the type of road emerges as the most influential factor in the model's performance. Specifically, higher road category values such as 'Vehicle track' and 'Motorway' on the left side exert a negative impact on the model, consistent with the observation that a majority of accidents occur in rural and urban areas. Moreover, Figure 2 demonstrates that drug consumption is associated with more severe accidents. The Shapley value linked to the drug-related feature increases proportionally with drug consumption. In essence, as levels of drug consumption escalate, so do the probabilities of accidents and the severity of resulting injuries.

TABLE 5. Results of MobileNet for traffic accident severity detection on US accidents.

Model	Accuracy	Precision	Recall	F1 score
MobileNet	0.9812	0.976	0.982	0.980

D. VALIDATION AND GENERALIZATION OF THE PROPOSED APPROACH

Besides validating using an alternative dataset, this study aimed to demonstrate the effectiveness of the proposed approach through experiments on a distinct dataset named the "US accidents (2016-2021)" dataset [41], which encompasses nearly 2.8 million records documenting traffic incidents occurring across 46 states in the United States between February 2016 and December 2021. This dataset provides extensive geographic coverage, spanning various locations across the country, and spans a five-year timeframe, making it valuable for analyzing regional and seasonal differences in accident patterns. It consists of 47 variables, detailing the causes contributing to automobile accidents, including information on accident locations, timings, weather conditions, road conditions, and accident severity.

The temporal intricacy in learning models primarily revolves around the instructional stage, gauging the duration necessary for adjusting the model's characteristics relying on the provided input information. This intricacy is impacted by various factors, encompassing the architectural intricacy of the model, the size of the instructional dataset, and the optimization method employed. Table 6 juxtaposes the durations for training and testing across the utilized models. Remarkably, the computational duration for the MobileNet model is exceptionally effective, enduring only 200 seconds, markedly less than the training periods of other transfer learning models employed in this investigation. Impressively, this increased efficiency in training time does not compromise the model's accuracy, as it consistently outperforms individual models in terms of predictive accuracy.

TABLE 6. Time complexity of transfer learning models (in seconds).

Model	Training Time	Testing Time
ResNET	123s	28s
EfficientNetB4	198s	52s
InceptionV3	150s	25s
Xception	110s	24s
VGG19	132s	28s
AlexNet	123s	21s
MobileNet	100s	19s

TABLE 7. Findings of k-fold cross-validation.

MobileNet	Accuracy	Precision	Recall	F-score
Fold-1	0.9873	0.9983	0.9873	0.9926
Fold-2	0.9797	0.9956	0.9895	0.9954
Fold-3	0.9879	0.9997	0.9994	0.9986
Fold-4	0.9962	0.9996	0.9996	0.9983
Fold-5	0.9947	0.9995	0.9993	0.9994
AVG.	**0.9891**	**0.9985**	**0.9950**	**0.9968**

To further evaluate the effectiveness of the proposed method, K-fold cross-validation is incorporated as an additional step for performance evaluation. The results from the 5-fold cross-validation are presented in Table 7. These findings showcase the exceptional efficacy of the suggested methodology in terms of precision, F1 score, accuracy, and recall when contrasted with alternative models. Notably, the low standard deviation values indicate consistent and stable performance across different folds, reinforcing confidence in the trustworthiness and reliability of MobileNet.

E. PRACTICAL IMPLICATIONS OF PROPOSED FRAMEWORK

The practical implications of utilizing the MobileNet transfer learning model and SHAP XAI technique for traffic accident severity prediction extend across various domains, including road safety management, emergency response, traffic control, insurance, fleet management, and public education. By leveraging these technologies, stakeholders can work towards reducing the frequency and severity of traffic accidents, ultimately saving lives and minimizing the societal and economic impact of road accidents.

1) Enhanced Road Safety Measures: By accurately predicting traffic accident severity, authorities can implement targeted road safety measures in high-risk areas. For instance, if the model predicts a high severity of accidents on a particular road segment, authorities can focus on improving road infrastructure, installing additional safety signs, or enforcing speed limits to mitigate the risk of accidents.
2) Optimized Emergency Response: Accurate predictions of accident severity enable emergency response teams to allocate resources more effectively. By anticipating the severity of accidents, emergency services can dispatch appropriate personnel and resources to the scene promptly, potentially reducing response times and improving outcomes for accident victims.
3) Improved Traffic Management: Traffic accident severity prediction can aid in optimizing traffic management strategies. Real-time information about potential accidents and their severity can help traffic management authorities reroute traffic, adjust signal timings, or implement temporary traffic controls to minimize congestion and maintain smooth traffic flow.
4) Insurance Risk Assessment: Insurance companies can leverage predictive models of accident severity to assess the risk associated with insuring vehicles and drivers. By accurately estimating the likelihood and severity of accidents, insurers can tailor insurance premiums more accurately, leading to fairer pricing for policyholders.
5) Fleet Safety Management: For businesses with vehicle fleets, predictive models of accident severity can inform fleet safety management strategies. By identifying routes or driving conditions associated with higher accident severity, fleet managers can implement training programs, safety protocols, or route optimizations to reduce the risk of accidents and protect both drivers and assets.
6) Public Awareness and Education: Transparent and explainable predictive models of traffic accident severity can also serve as educational tools for the general public. By providing insights into the factors influencing accident severity, such as road conditions, weather, or time of day, these models can raise awareness about safe driving practices and encourage responsible behavior among road users.

F. THEORETICAL CONTRIBUTIONS OF THE PROPOSED MODEL

The theoretical contributions of the research lie in the advancement of predictive modeling techniques, the integration of explainable AI methods, the validation of model interpretability, the understanding of feature importance, the application of transfer learning in road safety, and the demonstration of model generalization capabilities. These contributions collectively enhance the body of knowledge in the field of traffic accident severity prediction and pave the way for further advancements in AI-driven road safety solutions.

1) Advancement in Predictive Modeling Techniques: The research contributes to the advancement of predictive modeling techniques by integrating transfer learning with the MobileNet architecture. Transfer learning leverages knowledge gained from pre-trained models, enhancing the efficiency and effectiveness of predictive models for traffic accident severity prediction.
2) Integration of Explainable AI (XAI) Techniques: By incorporating SHAP (SHapley Additive exPlanations) XAI technique, the research enhances the interpretability and transparency of predictive models. This integration provides insights into the underlying factors

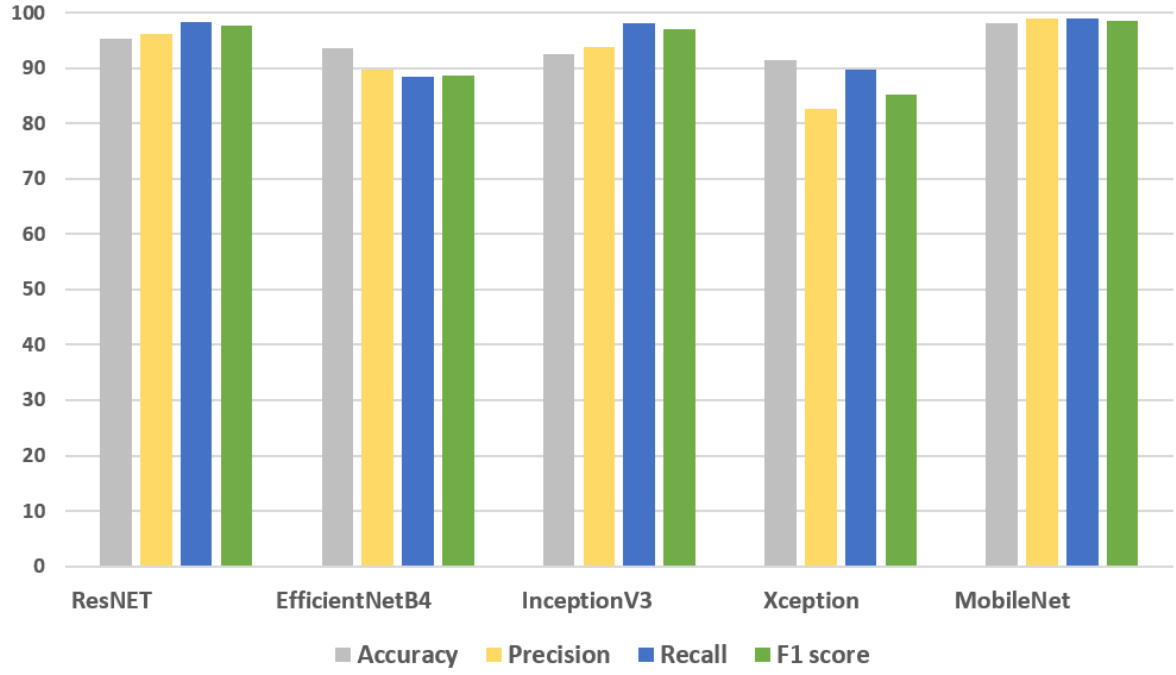

FIGURE 3. Transfer learning models performance comparison.

influencing accident severity predictions, contributing to the development of more transparent and trustworthy AI systems.

3) Validation of Model Interpretability: The research validates the effectiveness of SHAP XAI technique in interpreting complex machine learning models such as MobileNet. This contributes to the growing body of literature on model interpretability, demonstrating the applicability of SHAP in explaining predictions from deep learning models.
4) Understanding Feature Importance: Through the SHAP XAI technique, the research facilitates a deeper understanding of the importance of different features in predicting traffic accident severity. This understanding can inform future research on feature engineering, model optimization, and the development of more robust predictive models.
5) Application of Transfer Learning in Road Safety: The application of transfer learning in the context of road safety and accident severity prediction represents a novel contribution to the field. By leveraging pre-trained models, researchers can adapt state-of-the-art techniques from computer vision to address real-world problems in road safety management.
6) Demonstration of Model Generalization: The research demonstrates the generalization capabilities of the MobileNet transfer learning model across different datasets and real-world scenarios. This contributes to the understanding of model performance and robustness in the context of traffic accident severity prediction.

G. DISCUSSION

The discussion section undertakes a comprehensive evaluation of the study's findings and the implications of enhanced forecast accuracy for the severity of the accidents. The findings underscore the efficacy of deep learning approaches, particularly transfer learning models, in improving the exactness of predicting the severity of road traffic accidents. These findings carry significant implications for various stakeholders, including transportation planners, policymakers, and law enforcement agencies. The proposed MobileNet model can facilitate the implementation of targeted and efficient measures to reduce both the frequency and severity of traffic accidents by enabling more precise prediction of accident severity. Figure 3 illustrates the comparison of transfer learning models, clearly demonstrating the superiority of the proposed MobileNet.

Employing the MobileNet transfer learning model for predicting road accident severity offers numerous advantages. MobileNet arrives pre-trained with a wealth of knowledge from extensive image datasets, thereby reducing training time and mitigating the risk of overfitting. Its capability to generalize effectively to new and diverse accident images enhances predictive accuracy, while its efficient performance even on resource-constrained devices is noteworthy. MobileNet's consistent performance, resilience to image variations, and potential for fine-tuning render it a dependable

choice. Furthermore, its suitability for mobile deployment and potential to achieve state-of-the-art results make it a practical and high-accuracy solution for accident severity prediction.

Predicting accident severity holds significant real-world applications across various domains. It contributes to enhancing transportation safety, optimizing emergency response, assessing insurance risks, managing traffic, and improving fleet safety. Additionally, it plays a vital role in autonomous vehicle development, supports public health research, informs traffic engineering and urban planning initiatives, and aligns with smart city endeavors. Beyond road transportation, its applicability extends to aviation, maritime, and industrial safety, aiding in mitigating the impact of accidents, improving decision-making processes, and ultimately saving lives and resources.

1) COMPARISON WITH STATE-OF-THE-ART

For comparison purposes, two studies are selected. Yang et al. [16] applied the DNN model to detect injury severity and utilized the layer-wise relevance propagation (LRP) method to elucidate prediction outcomes. They conducted experiments using the Chinese traffic accident dataset. On the other hand, Cicek et al. [17] employed various models (DT, NB, MLP, SVM, NN, and ANN-MLP) and achieved 76.90% accuracy on the NHTSA-USA dataset. They utilized the Shapley technique to extract significant features. However, the NHTSA-USA dataset, while valuable for studying traffic accidents, presents limitations such as underreporting, inconsistent reporting, missing data, geographical and temporal biases, limited context, privacy concerns, data incompleteness, data imbalances, data collection bias, and evolving data standards. In contrast, the current study employs transfer learning models on the US traffic dataset and utilizes Shapley values to elucidate predictive outcomes. Shapley values, grounded in cooperative game theory, offer a more interpretable and versatile framework for explaining feature contributions in learning models. The current study demonstrates superior performance, achieving robust results with 98.17% accuracy.

V. CONCLUSION

Traffic accidents persist as a significant threat, resulting in loss of lives, injuries, and substantial disruptions on roadways. Understanding the underlying factors leading to these accidents is crucial for enhancing safety across transportation networks. This study harnesses various transfer learning techniques and elucidates the most influential factors through the application of Shapley values. The research delves into predicting accident severity using models including Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Residual Networks (ResNET), EfficientNetB4, InceptionV3, Extreme Inception (Xception), and MobileNet. Among these models, MobileNet emerges with the highest accuracy of 98.17%. This insight lays the groundwork for devising more effective measures to prevent accidents. By doing so, this research not only enhances the accuracy of severity prediction but also fosters transparency, interpretability, and trustworthiness of learning models. This is indispensable for stakeholders and decision-makers striving to implement evidence-based actions to bolster road safety, ultimately leading to a reduction in the impact of traffic accidents on highways. The future work direction of this research work is to make use of correlational and convergence feature engineering techniques to select the most appropriate factors that really affect the prediction of traffic accident severity in a real-time environment.

REFERENCES

[1] M. K. Gebru, "Road traffic accident: Human security perspective," *Int. J. Peace Develop. Stud.*, vol. 8, no. 2, pp. 15–24, 2017.

[2] N. Klinjun, M. Kelly, C. Praditsathaporn, and R. Petsirasan, "Identification of factors affecting road traffic injuries incidence and severity in Southern Thailand based on accident investigation reports," *Sustainability*, vol. 13, no. 22, p. 12467, 2021.

[3] *Road Safety*, World Health Organization, Geneva, Switzerland, 2020.

[4] A. Chand, S. Jayesh, and A. B. Bhasi, "Road traffic accidents: An overview of data sources, analysis techniques and contributing factors," *Mater. Today, Proc.*, vol. 47, pp. 5135–5141, Jan. 2021.

[5] F. Malin, I. Norros, and S. Innamaa, "Accident risk of road and weather conditions on different road types," *Accident Anal. Prevention*, vol. 122, pp. 181–188, Jan. 2019.

[6] V. Nuri Sumantri, A. I. Rifai, and F. Ferial, "Impact of inter-urban street lighting on users perception of road safety behavior: A case of jalan majalengka-rajagaluh," *Citizen, J. Ilmiah Multidisiplin Indonesia*, vol. 2, no. 5, pp. 703–711, Dec. 2022.

[7] C. Gutierrez-Osorio and C. Pedraza, "Modern data sources and techniques for analysis and forecast of road accidents: A review," *J. Traffic Transp. Eng.*, vol. 7, no. 4, pp. 432–446, Aug. 2020.

[8] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transp. Res. A, Policy Pract.*, vol. 44, no. 5, pp. 291–305, Jun. 2010.

[9] K. Santos, J. P. Dias, and C. Amado, "A literature review of machine learning algorithms for crash injury severity prediction," *J. Saf. Res.*, vol. 80, pp. 254–269, Feb. 2022.

[10] Y. Zhang, H. Li, and G. Ren, "Estimating heterogeneous treatment effects in road safety analysis using generalized random forests," *Accident Anal. Prevention*, vol. 165, Feb. 2022, Art. no. 106507.

[11] Y. Yang, K. He, Y.-P. Wang, Z.-Z. Yuan, Y.-H. Yin, and M.-Z. Guo, "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods," *Phys. A, Stat. Mech. Appl.*, vol. 595, Jun. 2022, Art. no. 127083.

[12] B. B. Elallid, N. Benamar, A. S. Hafid, T. Rachidi, and N. Mrani, "A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7366–7390, Oct. 2022.

[13] L. Li, Y. Lin, B. Du, F. Yang, and B. Ran, "Real-time traffic incident detection based on a hybrid deep learning model," *Transp. A, Transp. Sci.*, vol. 18, no. 1, pp. 78–98, 2022.

[14] K. Pawar and V. Attar, "Deep learning based detection and localization of road accidents from traffic surveillance videos," *ICT Exp.*, vol. 8, no. 3, pp. 379–387, Sep. 2022.

[15] H. Zhao, X. Li, H. Cheng, J. Zhang, Q. Wang, and H. Zhu, "Deep learning-based prediction of traffic accidents risk for Internet of Vehicles," *China Commun.*, vol. 19, no. 2, pp. 214–224, Feb. 2022.

[16] Z. Yang, W. Zhang, and J. Feng, "Predicting multiple types of traffic accident severity with explanations: A multi-task deep learning framework," *Saf. Sci.*, vol. 146, Feb. 2022, Art. no. 105522.

[17] E. Cicek, M. Akin, F. Uysal, and R. Topcu Aytas, "Comparison of traffic accident injury severity prediction models with explainable machine learning," *Transp. Lett.*, vol. 15, no. 9, pp. 1043–1054, Oct. 2023.

[18] J. Gan, L. Li, D. Zhang, Z. Yi, and Q. Xiang, "An alternative method for traffic accident severity prediction: Using deep forests algorithm," *J. Adv. Transp.*, vol. 2020, pp. 1–13, Dec. 2020.

[19] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, "Comparison of machine learning algorithms for predicting traffic accident severity," in *Proc. IEEE Jordan Int. Joint Conf. Electr. Eng. Inf. Technol. (JEEIT)*, Apr. 2019, pp. 272–276.
[20] I. Aldhari, M. Almoshaogeh, A. Jamal, F. Alharbi, M. Alinizzi, and H. Haider, "Severity prediction of highway crashes in Saudi Arabia using machine learning techniques," *Appl. Sci.*, vol. 13, no. 1, p. 233, Dec. 2022.
[21] M. Sameen and B. Pradhan, "Severity prediction of traffic accidents with recurrent neural networks," *Appl. Sci.*, vol. 7, no. 6, p. 476, Jun. 2017.
[22] L. Wahab and H. Jiang, "Severity prediction of motorcycle crashes with machine learning methods," *Int. J. Crashworthiness*, vol. 25, no. 5, pp. 485–492, Sep. 2020.
[23] D.-J. Lin, M.-Y. Chen, H.-S. Chiang, and P. K. Sharma, "Intelligent traffic accident prediction model for Internet of Vehicles with deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2340–2349, Mar. 2022.
[24] A. Jamal, M. Zahid, M. T. Rahman, H. M. Al-Ahmadi, M. Almoshaogeh, D. Farooq, and M. Ahmad, "Injury severity prediction of traffic crashes with ensemble machine learning techniques: A comparative study," *Int. J. Injury Control Saf. Promotion*, vol. 28, no. 4, pp. 408–427, Oct. 2021.
[25] M. Manzoor, M. Umer, S. Sadiq, A. Ishaq, S. Ullah, H. A. Madni, and C. Bisogni, "RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model," *IEEE Access*, vol. 9, pp. 128359–128371, 2021.
[26] T. K. Bahiru, V. S. Manjula, T. B. Akele, E. A. Tesfaw, and T. D. Belay, "Mining road traffic accident data for prediction of accident severity," in *Proc. Int. Conf. Intell. Data Commun. Technol. Internet Things (IDCIoT)*, Jan. 2023, pp. 606–612.
[27] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.
[28] M. T. C. Olmedo, M. Paegelow, J.-F. Mas, and F. Escobar, "Multilayer perceptron (MLP)," in *Geomatic Approaches for Modeling Land Change Scenarios*. Cham, Switzerland: Springer, 2018, pp. 451–455, doi: 10.1007/978-3-319-60801-3_27.
[29] J. Wu, "Introduction to convolutional neural networks," *Nat. Key Lab Novel Softw. Technol.*, vol. 5, no. 23, p. 495, 2017.
[30] R. C. Staudemeyer and E. Rothstein Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv:1909.09586*.
[31] F. He, T. Liu, and D. Tao, "Why ResNet works? Residuals generalize," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 12, pp. 5349–5362, Dec. 2020.
[32] R. Pillai, N. Sharma, and R. Gupta, "Fine-tuned EfficientNetB4 transfer learning model for weather classification," in *Proc. 3rd Asian Conf. Innov. Technol. (ASIANCON)*, Aug. 2023, pp. 1–6.
[33] N. Dong, L. Zhao, C. H. Wu, and J. F. Chang, "Inception v3 based cervical cell classification combined with artificially extracted features," *Appl. Soft Comput.*, vol. 93, Aug. 2020, Art. no. 106311.
[34] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[37] D. Sinha and M. El-Sharkawy, "Thin MobileNet: An enhanced MobileNet architecture," in *Proc. IEEE 10th Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Oct. 2019, pp. 0280–0285.
[38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
[39] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405.
[40] C. Yang, M. Chen, and Q. Yuan, "The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis," *Accident Anal. Prevention*, vol. 158, Aug. 2021, Art. no. 106153.
[41] Soobhan Moosavi. (May 2021). *US Accidents (2016–2021)*. Accessed: Feb. 10, 2023. [Online]. Available: https://www.kaggle.com/sobhanmoosavi/us-accidents

OMAR IBRAHIM ABOULOLA received the Bachelor of Science degree in computer science from KAU, in 2001, the master's degree in information science from the University of Indiana, Bloomington, USA, in 2009, and the master's degree in information technology and the Ph.D. degree in information systems and technology from Claremont Graduate University (CGU), USA, in 2013 and 2018, respectively. His master's thesis was related to the technology of banking. His Ph.D. dissertation aimed to design an innovative assistive technology to help retail companies to predict optimum locations for their businesses. He is currently an Associate Professor with the Information Systems and Technology Department, College of Computer Science and Engineering (CCSE), University of Jeddah.

EBTISAM ABDULLAH ALABDULQADER received the Ph.D. degree in HCI from Newcastle University, U.K. She is currently an experienced Assistant Professor with the Information Technology Department, King Saud University (KSU), Saudi Arabia. She is the Founder of the ArabHCI.org Community. Her research interests include social computing, participatory design, CSCW, and community-driven research.

AISHA AHMED ALARFAJ received the Ph.D. degree in computer science from Newcastle University, U.K., in 2021. Since 2012, she has been a Faculty Member with the Information Systems Department, Princess Nourah bint Abdulrahman University, where she is currently an Assistant Professor. Her research interests include human–computer interaction, UX/UI, social commerce, sharing economy, and trust.

SHTWAI ALSUBAI received the Ph.D. degree in computer science from The University of Manchester, London. He is currently an Assistant Professor with the Computer Engineering Department, Prince Sattam bin Abdulaziz University, Saudi Arabia. His research interests include computer vision, optimization techniques, and performance enhancement.

TAI-HOON KIM (Member, IEEE) received the M.S. and Ph.D. degrees in electrics, electronics, and computer engineering from Sungkyunkwan University, Seoul, South Korea, and the Ph.D. degree in information science from the University of Tasmania, Hobart, Australia, in December 2011. He is currently a Professor with Chonnam National University, Gwangju, South Korea. His research interests include statistical analysis, image processing, and system design.