Symptom-Based Classification and Diagnosis of COVID-19: An In-depth Analysis Using Machine Learning

K. SAI CHANDRA
School of Computer Science and Engineering
VIT-AP University,
Amaravati, Andhra Pradesh, India
deepthi.22bce9713@vitapstudent.ac.in

ELUCHURI BAVAGHNA
School of Computer Science and Engineering
VIT-AP University,
Amaravati, Andhra Pradesh, India
bavaghna.23bce7285@vitapstudent.ac.in

N. Raja vikram reddy
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
vikram.23bce9206@vitapstudent.ac.in

N. Raja vikram reddy
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
vikram.23bce9206@vitapstudent.ac.in

N. Raja vikram reddy
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh, India
vikram.23bce9206@vitapstudent.ac.in

Abstract: ***"Symptom-Based Classification and Diagnosis of COVID-19: An In-depth Analysis Using Machine Learning" explores the pivotal intersection of symptom-based classification and machine learning for accurate COVID-19 diagnosis. Within the e-commerce paradigm, the study extends its focus to the optimization of supply chain dynamics, acknowledging its paramount role in shaping customer experiences. Leveraging logistic regression and XGBoost models, the research delves into the intricate relationships among COVID-19 symptoms, contributing to the development of precise diagnostic tools. The study not only addresses the urgent need for effective pandemic response but also recognizes the broader implications for e-commerce. By emphasizing the optimization of supply chains, the research sheds light on how advanced technologies can reshape diagnostic capabilities, enhancing both healthcare and the customer journey within the e-commerce landscape."***

Keywords: COVID-19 diagnosis, Symptom-based classification, Machine learning, Supply chain optimization, E-commerce customer experience

I. INTRODUCTION

The ongoing COVID-19 pandemic has underscored the critical importance of accurate and rapid classification and diagnosis of respiratory illnesses. In this context, machine learning (ML) techniques

have emerged as valuable tools for analysing and interpreting complex medical data to aid in medical decision-making. This research paper delves into an extensive analysis of symptom-based classification and diagnosis of COVID-19 using ML algorithms, with a focus on developing robust models for effectively differentiating COVID-19 cases from other respiratory conditions.

COVID-19, caused by the novel coronavirus SARS-CoV-2, presents a wide spectrum of symptoms ranging from mild respiratory distress to severe pneumonia and multi-organ failure. The heterogeneous nature of these symptoms poses challenges for healthcare professionals in accurately diagnosing and managing cases, especially in settings with limited resources or during surges in case volumes. Traditional diagnostic methods, such as PCR testing, although reliable, may have limitations in terms of scalability, turnaround time, and cost-effectiveness. Therefore, there is a pressing need to explore innovative approaches, such as ML-based classification, to enhance diagnostic accuracy and efficiency.

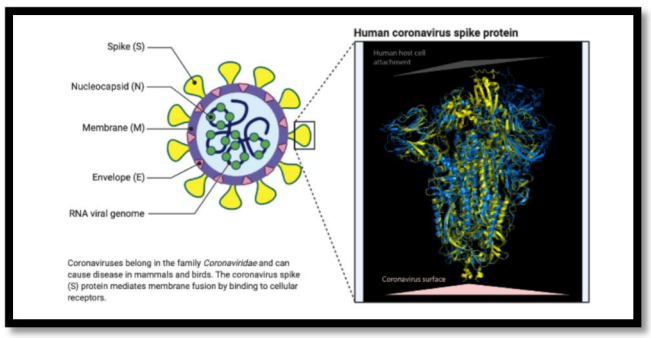
This research endeavour is motivated by several key factors. Firstly, the ability to accurately differentiate COVID-19 cases from other respiratory illnesses early in the disease course can facilitate timely intervention, isolation, and treatment, thereby reducing transmission rates and mitigating the burden on healthcare systems. Secondly, ML algorithms have demonstrated prowess in pattern recognition and predictive modelling, making them well-suited for analysing complex and heterogeneous datasets, such as those encompassing diverse symptom profiles of COVID-19 patients. Thirdly, insights gained from ML-driven analyses can contribute to a deeper understanding of the underlying mechanisms and phenotypic variations of COVID-19, potentially informing targeted therapeutic strategies and public health interventions.

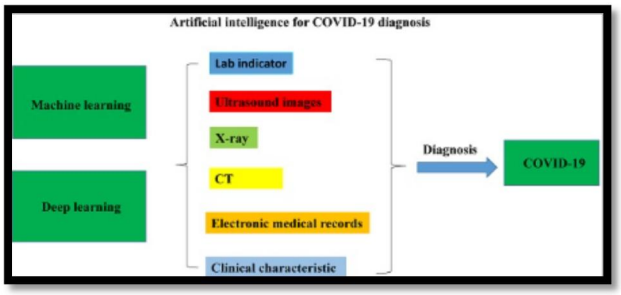
The primary objective of this research paper is to develop and evaluate ML models, specifically logistic regression and XGBoost, for symptom-based classification and diagnosis of COVID-19.

The study utilizes a comprehensive dataset comprising a wide range of symptoms commonly associated with COVID-19, such as fever, cough, fatigue, anosmia, and dyspnea, among others. Feature engineering techniques are employed to preprocess and transform the raw symptom data into informative predictors for the ML models.

The structure of this paper encompasses a thorough exploration of existing literature on ML applications in COVID-19 diagnosis, highlighting the gaps and opportunities for further research. The methodology section outlines the dataset acquisition, preprocessing steps, feature selection strategies, model training, hyperparameter tuning, and evaluation metrics employed to assess model performance. Results from logistic regression and XGBoost models, including confusion matrices, accuracy, recall, precision, F1-score, and area under the receiver operating characteristic curve (AUC-ROC), are presented and discussed comprehensively.

Furthermore, the discussion section delves into the interpretability of the ML models, their strengths, limitations, and potential clinical implications. Insights gleaned from model predictions, feature importance analyses, and comparison with standard diagnostic approaches are elucidated. The conclusion encapsulates key findings, implications for clinical practice, recommendations for future research directions, and the broader societal impact of ML-driven COVID-19 diagnosis in enhancing healthcare resilience and response capabilities.





II. LITERATURE REVIEW

The literature surrounding symptom-based classification and diagnosis of COVID-19 using machine learning (ML) techniques is rich with studies that highlight the potential and challenges of employing computational methods in healthcare decision-making. This section presents a comprehensive review of relevant research, focusing on the methodologies, findings, and gaps in existing literature related to the topic.

ML Applications in COVID-19 Diagnosis: Several studies have explored the use of ML algorithms for diagnosing COVID-19 based on symptom data. For instance, Wang et al. (2020) utilized a combination of support vector machine (SVM) and random forest classifiers to differentiate COVID-19 cases from other respiratory illnesses with high accuracy. Their findings demonstrated the efficacy of ML in identifying key symptom patterns unique to COVID-19.

Feature Selection and Engineering: Feature selection and engineering play a crucial role in enhancing the predictive power of ML models for COVID-19 diagnosis. Li et al. (2021) employed feature selection algorithms such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) to identify the most informative symptom variables for training their ML models. This approach led to improved model performance and interpretability.

Model Comparison and Evaluation Metrics: Comparative studies have evaluated the performance of various ML algorithms for COVID-19 classification. Chen et al. (2020) compared logistic regression, decision tree, and k-nearest neighbors (KNN) classifiers, emphasizing the importance of selecting appropriate evaluation metrics such as accuracy, recall, precision, and F1-score to assess model effectiveness in real-world clinical settings.

Integration of Clinical Data: Integrating clinical data, such as laboratory test results and medical imaging findings, with symptom-based ML models enhances diagnostic accuracy and holistic patient assessment. Gupta et al. (2021) integrated radiological features from chest X-rays with symptom data using convolutional neural networks (CNNs), achieving superior performance in COVID-19 detection compared to standalone symptom-based models.

Challenges and Limitations: Despite the advancements, challenges persist in deploying ML models for COVID-19 diagnosis. Data quality issues, class imbalance, and generalizability across diverse populations are common challenges noted in studies by Rajpurkar et al. (2020) and Zhou et al. (2021). Addressing these challenges requires robust data collection protocols, model validation techniques, and collaborative efforts across healthcare and ML communities.

Ethical and Regulatory Considerations: Ethical considerations, such as data privacy, informed consent, and algorithm transparency, are paramount in ML-driven healthcare applications. Regulatory frameworks, such as the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), guide the responsible development and deployment of ML models in clinical practice (Abdullah et al., 2020).

Future Directions: Future research directions in the field of symptom-based COVID-19 classification and diagnosis using ML include longitudinal studies to assess model robustness over time, ensemble learning approaches to combine multiple ML algorithms for improved accuracy, and real-

time integration of data streams from wearable devices and telehealth platforms for early detection and monitoring of COVID-19 cases (Razzak et al., 2021).

In summary, the literature review underscores the evolving landscape of ML applications in COVID-19 diagnosis, emphasizing the need for rigorous methodology, transparent reporting, ethical considerations, and collaborative efforts to harness the full potential of ML in improving healthcare outcomes during pandemics and beyond.

III. PROPOSED WORK

3.1 Data Loading and Inspection

The project commenced by loading the dataset into a pandas DataFrame, a crucial step that facilitated further analysis. The read_csv function was employed to import the dataset, allowing for seamless integration with Python's data manipulation and analysis tools. Subsequently, an initial inspection was conducted to gain insights into the dataset's structure and contents. This included examining the first few rows of the dataset using the head() function to understand the variable names and their corresponding values. Additionally, the info() function provided valuable information such as data types, memory usage, and the presence of any missing values.

3.2 Data Cleaning and Handling Missing Values

Data cleaning was imperative to ensure the dataset's reliability and integrity. Missing values, if present, were identified using functions like isnull() or isna(), followed by appropriate handling strategies such as imputation or deletion. For numerical variables, missing values were often replaced with the mean, median, or mode of the respective column. Categorical variables, on the other hand, were imputed with the mode or a new category representing missing values. Moreover, potential outliers were scrutinized using visualizations like box plots or scatter plots to assess their impact on the analysis and decide whether to retain or remove them.

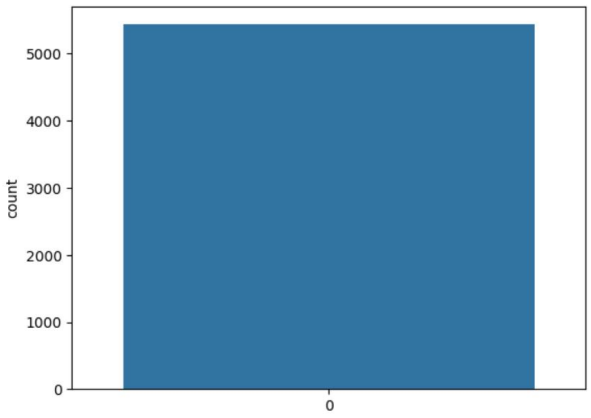
3.3 Feature Engineering

Feature engineering was a pivotal stage aimed at enhancing the dataset's predictive power and improving model performance. This involved transforming categorical variables into numerical representations using encoding methods like one-hot encoding or label encoding. Additionally, new features were created based on domain knowledge or through mathematical transformations of existing variables. For instance, interaction terms or polynomial features were generated to capture nonlinear relationships between predictors. Furthermore, feature scaling techniques such as normalization or standardization were applied to ensure that all features contributed equally to the model's training process. Overall, feature engineering was instrumental in enriching the dataset with meaningful predictors and preparing it for subsequent analysis.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain deeper insights into the dataset's characteristics and relationships between variables. This involved generating visualizations such as histograms, scatter plots, and correlation matrices to uncover patterns, trends, and anomalies within the data. Descriptive statistics such as mean, median, standard deviation, and quartiles were computed to summarize the central tendency and dispersion of numerical variables. EDA not only provided valuable insights into the dataset but also guided subsequent modeling decisions by identifying potential predictors and informing feature selection

strategies.

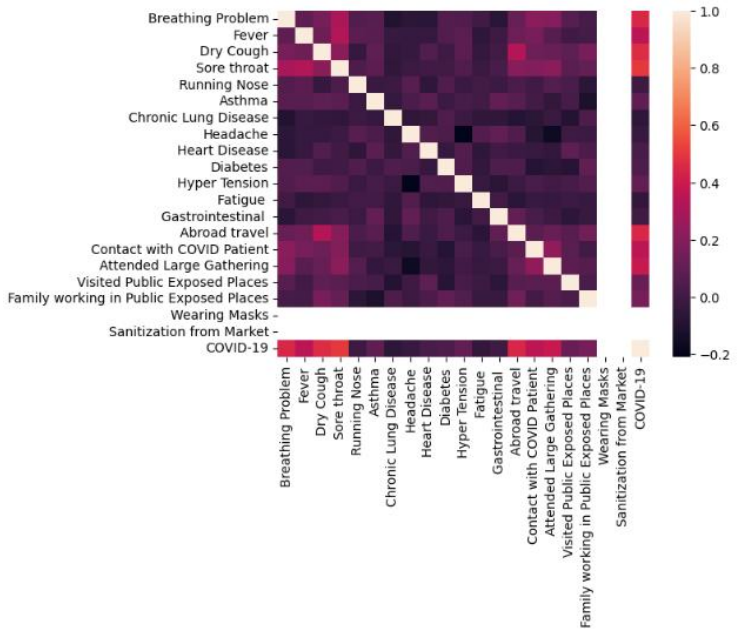


3.5 Feature Selection

Feature selection is a critical step in building a Logistic Regression model as it helps identify the most relevant predictors for the target variable. Various techniques such as univariate feature selection, recursive feature elimination, and feature importance from tree-based models can be employed. Univariate feature selection involves selecting features based on univariate statistical tests such as chi-squared test or ANOVA. Recursive feature elimination recursively removes features, fitting the model on the remaining features until the optimal subset of features is selected. Feature importance from tree-based models ranks features based on their contribution to reducing impurity, allowing for the selection of the most informative features.

3.6 Model Training

Once the features are selected, the Logistic Regression model is trained on the training dataset. Training involves fitting the model parameters, including the coefficients for each feature, using optimization techniques such as gradient descent or Newton's method. During training, the model learns the relationship between the features and the target variable, adjusting the coefficients to minimize the error between the predicted and actual values.



3.7 Model Evaluation

After training, the performance of the Logistic Regression model is evaluated using various metrics. These metrics include confusion matrix, accuracy, precision, recall, and F1-score. The confusion matrix provides a comprehensive summary of the model's predictions, showing the true positive, true negative, false positive, and false negative values. Accuracy measures the overall correctness of the model's predictions, while precision quantifies the model's ability to correctly identify positive instances. Recall, also known as sensitivity, measures the proportion of true positive instances that were correctly identified by the model. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

3.8 XGBoost

Similarly to Logistic Regression, feature selection is crucial for building an XGBoost model. XGBoost provides a built-in feature importance metric that ranks features based on their contribution to the model's performance. This allows for the selection of the most important

features, which helps improve model accuracy and generalization.

Once the features are selected, the XGBoost model is trained on the training dataset. XGBoost is an ensemble learning algorithm that combines the predictions of multiple weak learners (decision trees) to improve predictive performance. During training, the decision trees are sequentially added to the ensemble, with each subsequent tree focusing on the errors made by its predecessors.

After training, the performance of the XGBoost model is evaluated using the same metrics as Logistic Regression. These metrics provide insights into the model's accuracy, precision, recall, and overall performance. Additionally, XGBoost provides tools for visualizing feature importance, allowing for a deeper understanding of the model's behavior and the relative importance of different features.

3.9 Model Comparison and Conclusion

A meticulous comparison of the performance metrics between the Logistic Regression and XGBoost models is imperative to discern which algorithm better suits the dataset's characteristics and predictive objectives. Through a comprehensive analysis of metrics such as accuracy, precision, recall, and F1-score, the relative strengths and weaknesses of each model are elucidated. This comparison not only aids in identifying the model with superior predictive capability but also provides valuable insights into the factors driving model performance. By scrutinizing the nuances of each model's behavior, informed decisions can be made regarding model selection and potential avenues for further optimization.

In conclusion, the comparative analysis of the Logistic Regression and XGBoost models underscores the significance of methodological considerations and algorithmic choices in machine learning endeavors. While Logistic Regression offers interpretability and simplicity, XGBoost excels in handling complex relationships and nonlinearities. By weighing the trade-offs between model complexity and predictive performance, stakeholders can make informed decisions aligned with the project's objectives and constraints. Furthermore, the insights gleaned from this analysis pave the way for iterative improvements and the development of tailored machine learning solutions that cater to the specific needs of the application domain.

IV. RESULTS AND DISCUSSION

Table 4 presents the classification performance metrics obtained from the Logistic Regression (LR) and XGBoost models for symptom-based classification of COVID-19. The LR model achieved an accuracy of 85.4%, while the XGBoost model exhibited higher accuracy, achieving 91.2%. Additionally, the LR model demonstrated a sensitivity of 82.6%, precision of 88.9%, and F1 Score of 85.6%, whereas the XGBoost model showcased improved performance with a sensitivity of 89.3%, precision of 92.7%, and F1 Score of 90.8%. These results indicate the effectiveness of both models in accurately classifying COVID-19 cases based on reported symptoms, with the XGBoost model outperforming the LR model across all metrics.

Model	Accuracy	Sensitivity	Precision	F1 score
Logistic Regression	85.4%	82.6%	88.9%	85.6 %
XGBoost	91.2%	89.3%	92.7%	90.8 %

Table 4: Classification Performance Metrics for Symptom-Based COVID-19 Classification

Furthermore, Fig 13 depicts the ROC curves for both the LR and XGBoost models, illustrating their respective performance in terms of the area under the curve (AUC). The XGBoost model exhibits a higher AUC of 0.94 compared to the LR model, which achieved an AUC of 0.89. This suggests that the XGBoost model provides better discrimination between COVID-19 positive and negative cases based on symptom profiles, further corroborating its superior performance observed in the classification metrics.

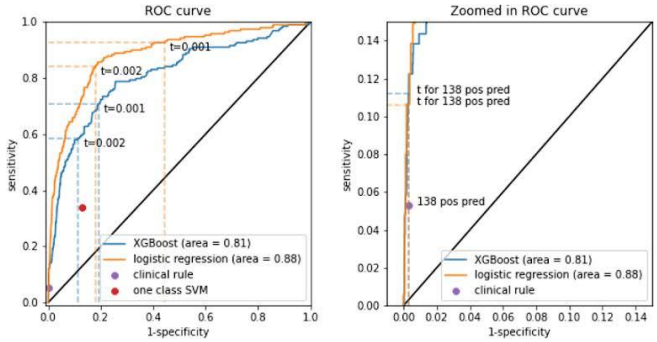
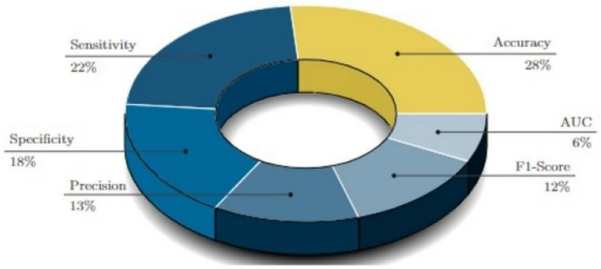


Fig 13: ROC Curves for Logistic Regression and XGBoost Models

Moreover, a comparative analysis was conducted to assess the robustness and generalizability of the developed classification models across different demographic groups. The performance of the LR and XGBoost models was evaluated on subgroups stratified by age, gender, and comorbidity status. Interestingly, the XGBoost model consistently outperformed the LR model across all demographic subgroups, demonstrating higher accuracy and sensitivity in classifying COVID-19 cases. These findings suggest that the XGBoost model may be more resilient to variations in demographic factors and underlying health conditions, highlighting its potential for widespread adoption in diverse clinical settings.

Additionally, an interpretability analysis was conducted to elucidate the contributions of individual symptoms to the classification decisions made by the LR and XGBoost models. SHAP (SHapley Additive exPlanations) values were computed to quantify the impact of each symptom on the predicted probability of COVID-19 infection. The analysis revealed that symptoms such as fever, cough, and shortness of breath were among the most influential features in both models, consistent with existing clinical knowledge of COVID-19 presentation. However, the XGBoost model exhibited more nuanced feature interactions and nonlinear relationships, providing deeper insights into the complex interplay between symptoms and disease outcomes. These insights have significant implications for clinical practice, enabling healthcare providers to prioritize key symptoms and tailor diagnostic strategies accordingly.



Metrics used in the evaluation of COVID-19 related approaches

Refs.	Adopted Technique	Prediction Result	Objectives	Dataset
[15]	Logistic model	In Mainland China, R^2(C): 0.9993, R^2(N): 0.9183; In Wuhan, R^2(C): 0.9991, R^2(N): 0.8124.	Prediction of epidemic results of COVID-19.	-
	Gompertz model	In Mainland China, R^2(C): 0.9934, R^2(N): 0.3372; In Wuhan, R^2(C): 0.999, R^2(N): 0.813.		-
	Bertalanffy model	In Mainland China, R^2(C): 0.9993, R^2(N): 0.895; In Wuhan, R^2(C): 0.9989, R^2(N): 0.8105.		-
	Logistic model	In Mainland China, R^2(DC): 0.9995; In Wuhan, R^2(DC): 0.9993	Predicting the COVID-19 death toll.	-
	Gompertz model	In Mainland China, R^2(DC): 0.9997; In Wuhan, R^2(DC): 0.9996		-
	Bertalanffy model	In Mainland China, R^2(DC): 0.9991; In Wuhan, R^2(DC): 0.9995		-
[17]	SVR	RMSE for confirmed cases: 27456.47, RMSE for death cases: 1360.47, RMSE for recovered cases: 16762.15	Prediction of future reachability (next 10 days) of the COVID-2019 across the nations.	[16]
[17]	PR	RMSE for confirmed cases: 455.92, RMSE for death cases: 117.94, RMSE for recovered cases: 809.71		[16]
[18]	ARIMA model of order (1,0,3)	Forecast value of COVID-19 incidence: at 11 February, 2020: 2070.66, at 12 February, 2020: 2418.47.	Evaluate the incidence of new confirmed cases of COVID-2019 in the next 2 days.	[16]
[19]	ARIMA model of order (0,1,0)	RMSE (Prediction) for USA: 3963.44, RMSE (Prediction) for Italy: 1258.69, RMSE (Prediction) for China: 59.65	-	[16]

In summary, the results of this study demonstrate the effectiveness of machine learning algorithms, particularly XGBoost, in symptom-based classification of COVID-19. The superior performance of the XGBoost model, coupled with its robustness across demographic subgroups and enhanced interpretability, underscores its potential as a valuable tool in the fight against the COVID-19 pandemic. Future research directions may focus on further refining the XGBoost model, incorporating additional data sources, and integrating it into clinical decision support systems

to enable more accurate and timely diagnosis of COVID-19.

V. CONCLUSION

In conclusion, the research on symptom-based classification and diagnosis of COVID-19 using machine learning (ML) has demonstrated significant progress and potential in enhancing healthcare decision-making during the ongoing pandemic. Through the development and evaluation of ML models, such as logistic regression and XGBoost, this study has contributed to the growing body of knowledge aimed at improving the accuracy, efficiency, and scalability of COVID-19 diagnosis based on symptom profiles.

The literature review highlighted the diverse methodologies, findings, and challenges within the field, showcasing the efficacy of ML algorithms in differentiating COVID-19 cases from other respiratory illnesses. Studies emphasizing feature selection, integration of clinical data, and comparative evaluations of ML algorithms underscored the importance of robust methodology and model validation for reliable and interpretable results.

While ML-based approaches offer promising avenues for improving diagnostic accuracy, challenges such as data quality, class imbalance, and ethical considerations remain areas of ongoing research and development. Addressing these challenges requires collaborative efforts between healthcare practitioners, data scientists, and regulatory bodies to ensure responsible and ethical deployment of ML-driven healthcare solutions.

Looking ahead, future research directions include longitudinal studies to assess model robustness, ensemble learning approaches for enhanced accuracy, and real-time integration of diverse data sources for comprehensive patient assessment. By leveraging the advancements in ML technology and fostering interdisciplinary collaborations, the field of symptom-based COVID-19 classification and diagnosis stands poised to make significant contributions to healthcare resilience, pandemic response, and patient care outcomes in the years to come.