Computers and Electrical Engineering 118 (2024) 109311

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

IoT based smart framework to predict air quality in congested traffic areas using SV-CNN ensemble and KNN imputation model

Khaled Alnowaiser [a], Aisha Ahmed Alarfaj [b], Ebtisam Abdullah Alabdulqader [c], Muhammad Umer [d], Lucia Cascone [e,*], Bhavya Alankar [f]

[a] Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, P.O. Box 151, 11942, Saudi Arabia
[b] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[c] Department of Information Technology, College of Computer and Information Sciences, King Saud University, Saudi Arabia
[d] Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan
[e] Department of Computer Science, University of Salerno, Fisciano, Italy
[f] Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi 110062, India

ARTICLE INFO

Keywords:
Air quality prediction
IoT & air quality control
Smart city pollution control
KNN Imputer
Ensemble learning

ABSTRACT

Addressing air pollution presents a significant environmental challenge in the context of smart city environments. Real-time monitoring of pollution data empowers local authorities to assess current traffic conditions and make informed decisions accordingly. The integration of Internet of Things (IoT) sensors has revolutionized air quality prediction, with human activities being the primary contributors to air pollution, posing threats to all forms of life. Gases such as SO2, PM10, O3, CO, NO2, among others, are key pollutants. Exposure to air pollution can lead to severe health issues and fatalities, underscoring the criticality of air quality monitoring. Distinguishing between breathable and non-breathable air quality further enhances the value of air quality monitoring. However, existing techniques face challenges in achieving high accuracy, exacerbated by missing values in datasets, which can significantly impact machine learning model performance. In response to these challenges, this research introduces an IoT-based automated system designed to classify air quality. This system is adept at managing missing data and attaining high accuracy levels. Central to this approach is a stacked ensemble voting classifier model, which combines two machine learning models. Additionally, the system incorporates the KNN Imputer to address missing values effectively. The work assesses the system's performance against seven alternative machine learning algorithms across two scenarios: one with missing values removed and another with KNN imputation applied. Notably, the proposed strategy achieves remarkable metrics, including 99.17% accuracy, 97.75% precision, 95.24% recall, and a 96.52% F1 score, when leveraging the KNN Imputer. These results underscore the effectiveness of the proposed model compared to current state-of-the-art methodologies.

1. Introduction

Air is essential for human survival, and its quality is crucial to our well-being. However, air pollution has become a major concern worldwide, posing significant risks to all living organisms, including plants, animals, and humans. According to the World

* Corresponding author.
E-mail addresses: k.alnowaiser@psau.edu.sa (K. Alnowaiser), Aiaalarfaj@pnu.edu.sa (A.A. Alarfaj), eabdulqader@ksu.edu.sa (E.A. Alabdulqader), umersabir1996@gmail.com (M. Umer), lcascone@unisa.it (L. Cascone), bhavya.alankar@gmail.com (B. Alankar).

https://doi.org/10.1016/j.compeleceng.2024.109311
Received 19 February 2024; Received in revised form 18 April 2024; Accepted 14 May 2024
Available online 31 May 2024

Health Organization (WHO), 9/10 persons are breathing polluted air, leading to 7 million deaths annually [1]. Environmental impacts of poor air quality include acid rain, smog, global warming, climate change, reduced visibility, and genetic mutations. Scientific evidence shows that air pollution adversely affects historical monuments [2]. Emissions of greenhouse gases from vehicles, agricultural activities, and industries negatively influence climate conditions and plant growth [3]. Additionally, these emissions impact plant-soil interactions [4] and degrade agricultural productivity and quality, potentially causing economic losses and famine [5].

Economic growth and environmental protection are often viewed as competing goals, particularly in developing countries [6]. For these reasons, the regulatory frameworks and enforcement mechanisms for pollution control vary significantly across different countries and regions due to variations in economic development, environmental priorities, and political systems. In many regions, the immediate imperative of economic survival outweighs environmental concerns. Communities that rely on polluting industries for employment may oppose pollution control measures that could jeopardize jobs. However, concurrently with escalating environmental degradation and climate change, a renewed public consciousness regarding environmental preservation has emerged. Nonetheless, barriers to behavioral change continue to exist within both the physical and social environments, and cannot be overcome. A primary barrier is the lack of accessible and reliable information on pollution levels and their health impacts [7]. Moreover, for example, the decision to cycle or use public transport often depends on factors such as travel distances, the accessibility of affordable public transportation, or the availability of bicycle paths. Therefore, despite concerted efforts by nations worldwide to address this pressing issue through policy interventions and strategies, there remains a notable deficit in public awareness, clear guidelines, and effective methodologies for pollution mitigation [8]. Consequently, pollution rates have continued to rise steadily over recent decades. While numerous studies have shed light on the causes, contributing factors, and potential mitigation strategies for air pollution, achieving comprehensive eradication at local, regional, and global scales remains an ongoing challenge [9–11].

Key atmospheric pollutants such as particulate matter (PM), nitrogen dioxide (NO2), sulfur dioxide (SO2), ozone (O3), and carbon monoxide (CO) have significant impacts on human health, ecosystems, and climate. PM, emitted from vehicular exhausts and industrial processes, can penetrate deep into the lungs, causing respiratory illnesses and cardiovascular problems. NO2, primarily from vehicle emissions and industrial combustion, exacerbates respiratory conditions and contributes to the formation of smog and acid rain, harming ecosystems and infrastructure. SO2 emissions from fossil fuel combustion lead to acid rain, affecting soil quality and aquatic ecosystems, while also contributing to respiratory issues in humans. Ozone, a secondary pollutant formed from NO2 and volatile organic compounds (VOCs) in the presence of sunlight, damages lung tissue and reduces crop yields, impacting human health and agricultural productivity. Additionally, CO, primarily from vehicle emissions and incomplete combustion, impairs oxygen delivery in the bloodstream, posing health risks, and contributes to the greenhouse effect, exacerbating climate change [12]. These pollutants underscore the interconnectedness of atmospheric pollution with human health, ecosystem integrity, and climate stability. The dataset, we have utilized in this research work contains most of these key elements to predict the air-quality in congested areas. The air quality index (AQI) is a parameter used to assess air quality, and it is directly related to public health. Air pollution level is numerically communicated and measured as AQI. CO, NO2, O3, SO2, particles having a size less than 10 μm (PM10), particles having a size less than 2.5 μm (PM2.5), benzene, and NH3 are among the 12 AQI parameters and are accounted for as prominent air pollutants. Sometimes, PM2.5, SO2, CO, NO2, O3, and PM10 are the only 6 factors used to measure AQI. Pollution selection depends upon the data availability, monitoring frequency, measurement techniques, and specific purposes. A high AQI refers to highly polluted air, which can have adverse impacts on health.

Rapid industrialization and population growth have resulted in increased levels of toxic gas emissions, severely impacting human health. Unchecked pollution has led to a significant decline in air quality, and it is important that we monitor and understand air quality to prevent further harm. The increasing population contributes to a rise in air pollution, with more pollutants being released into the atmosphere each year. This poor air quality index is life-threatening for human beings. Air pollution can be broadly classified into two main types — outdoor and indoor air pollution. While both of these types are hazardous on their own, they can also worsen each other's effects since air can circulate between indoor and outdoor environments. As a result, indoor air pollution can affect the quality of outdoor air and vice versa [13].

The percentage of the world's urban population is on the rise as more and more people are moving to cities. According to the United Nations (UN), as of 2020, the urban population is approximately 56.15% [14]. Furthermore, the urban population worldwide is expected to grow to a whopping 68% by 2050 [15]. The growth of urbanization and industrialization has given rise to various challenges such as logistics, healthcare, and air pollution. To tackle these issues and improve the quality of life for its citizens, the concept of smart cities has emerged. This involves integrating information and communication technology (ICT) and installing fixed/mobile sensors throughout the city to monitor real-life human activities, generating a wealth of urban data. However, monitoring and measuring air pollution levels poses significant challenges, particularly in densely populated urban areas and remote regions. In urban areas, the high density of sources such as vehicular emissions, industrial activities, and residential heating makes it difficult to accurately measure and attribute pollutant concentrations. Additionally, the complex and dynamic nature of urban airflows can lead to spatial and temporal variations in pollution levels, requiring a dense network of monitoring stations for comprehensive coverage. Furthermore, the cost and maintenance of monitoring equipment, as well as the need for skilled personnel to operate and analyze data, present logistical challenges. In remote regions, access to infrastructure and resources for monitoring air quality is limited, making it challenging to obtain representative data. Additionally, factors such as terrain, weather conditions, and seasonal variations can further complicate measurements in remote areas. Addressing these challenges requires innovative monitoring technologies, improved data-sharing mechanisms, and collaborative efforts among government agencies, research institutions, and communities to ensure effective air quality management [16].

Table 1
Overview of previous studies in air quality prediction.

Reference	Year	Dataset	Classifiers	Accuracy
[21]	2020	Dataset maintained by the Ministry of Environment, Jordan	Decision Tree (DT), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF)	92.01%
[22]	2020	US Environmental Protection Agency (US EPA) dataset	Support Vector Regression (SVR) with various feature selection techniques	94.1%
[23]	2021	Datasets comprising pollutant concentration and meteorological factors	Artificial Neural Network (ANN), Random Forest (RF), Support Vector Machine (SVM)	99.40%
[24]	2019	Central Pollution Control Board (CPCB), India Dataset	Neural Networks (NNs)	95.01%
[25]	2022	CPCB, India Dataset	Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost)	90.01%
[26]	2019	CPCB, India Dataset	Support Vector Machine (SVM), Neural Networks (NNs)	97.30%
[27]	2021	CPCB, India Dataset	Synthetic Minority Over-sampling Technique (SMOTE)-based Deep Neural Network (DNN), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN)	90.90%
[25]	2023	CPCB, India Dataset	Support Vector Regression (SVR), Random Forest Regression (RFR), Classification and Regression (CR)	97.60%
[28]	2022	Malaysia's Ministry of Environment and Water's Department of Environment (DoE) Dataset	Decision Tree (DT), Boosted Regression Tree (BRT), Random Forest (RF)	98.30%

Machine learning and deep learning models can be used for automated air quality prediction and have been used for the prediction of various atmospheric elements [17,18]. It falls under the domain of artificial intelligence (AI) on the idea that informed decisions without human intervention can be made by identifying the data patterns based on system learning from available data. Before such an application, the system is trained on various related datasets. As a result, machine learning is utilized in nearly all areas of science and technology in modern times [19,20]. Keeping in view the potential of machine learning models, this study presents a machine learning approach for air quality prediction and makes the following contributions:

- This study introduces a novel ensemble model termed SV-CNN for air quality prediction in environmental contexts. SV-CNN combines support vector machine (SVM) and convolutional neural network (CNN) architectures, with final predictions made via soft voting.
- Given potential missing values in the data, which can impact model performance, this study employs the K nearest neighbor (KNN) Imputer to address this issue. An essential aspect involves conducting experiments both with and without the KNN Imputer to assess its impact on results.
- The study conducts a comprehensive performance comparison with state-of-the-art models including random forest (RF), logistic regression (LR), gradient boosting machine (GBM), extra tree classifier (ETC), SVM, decision tree (DT), and stochastic gradient descent (SGD). Additionally, the effectiveness of the proposed model is evaluated by comparing its performance metrics – accuracy, precision, recall, and F1 score – with those of state-of-the-art approaches.

The paper's organization is structured as follows: Section 2 provides a concise overview of prior research in this domain. Section 3 delineates the dataset, the machine learning models employed for air quality prediction, and the proposed methodology, offering detailed insights into the approach. Section 4 delves into the discussion of results. Finally, Section 5 presents the conclusions drawn based on the findings.

2. Related work

Clean and clear air is crucial to the survival of both humans and plants, and its availability is necessary for human existence. However, air quality is often impacted by various pollutants, to varying degrees. Due to the importance of clean air in human life, an adaptable, accurate, and trustworthy air pollution prediction model is needed. Consequently, air quality assessment has gained significant interest. Recently, air quality has been precisely predicted using machine learning algorithms and neural networks by various researchers. However, these models have failed to establish and generalize highly non-linear and complex connections between modeling parameters.

Castelli et al. [22] focused on the prediction of air quality in the US in the state of California by projecting pollutant levels using the support vector regressor (SVR) algorithm. An innovative approach has been proposed by the authors to model atmospheric pollution on an hourly basis. Nahar et al. [21] created a model that utilizes machine learning classifiers to predict AQI. The authors analyzed data collected by Jordanian Ministry of Environment, over a period of 28 months and noticed the concentrations of pollutants. This model was able to accurately detect the most polluted areas. Soundari et al. [24] created a model using neural

networks to predict the Indian AQI. The authors found that they can predict the AQI of an entire country, or any other geographical location, based on their anticipated model while a dataset related to air pollutants' concentrations is available.

The authors conducted a study on a dataset containing information on both the pollutants' concentration and climatological factors [23]. It is found that the RF classifier outperformed other methods due to its ability to reduce overfitting. A model was created by Mahalingam et al. in [26] to anticipate the air quality index of smart cities. It was concluded that the medium Gaussian SVM achieved the highest precision. The suggested model could be utilized in any smart city according to the authors. Kumar & Pande [25] utilized the dataset used in this study which had already undergone preprocessing, including the selection of key features through correlation analysis. The authors conducted exploratory data analysis to explore hidden patterns in the dataset and identified the pollutants directly linked to the AQI. The study also observed a considerable decline in the concentration of almost all pollutants during 2020 due to the pandemic. To address the data imbalance issue, AQI was predicted by the authors using five models and deploying the resampling technique. The XGBoost model was found to outperform the rest of the models, demonstrating the highest linearity between the actual and the forecasted data.

In another study [25], the authors conducted a comparative analysis of three machine learning algorithms CatBoost regression (CR), RF regression (RFR), and SVR for predicting AQI in four major cities of India. The authors utilized the synthetic minority oversampling technique (SMOTE) to improve the accuracy of the models. The results showed that RFR outperformed the other models and achieved an accuracy score of 97.6080% on the data using SMOTE. The study was conducted with intensive research to obtain the most feasible solution for air pollution. In [27], the authors proposed a machine learning system that employed the SMOTEDNN oversampling technique and five different models. The system included effective data pre-processing and rigorous hyper-parameter optimization. The proposed system achieved an impressive accuracy score of 99.90%. However, it should be noted that the system requires careful hyperparameter tuning, which can be a time-consuming process.

Shaziayani and colleagues [28] introduced an innovative methodology to forecast and categorize PM10 concentrations employing tree-based machine learning techniques, including boosted regression trees, decision trees (DT), and random forests (RF) within the Malaysian context. Particularly noteworthy was the study's focus on predicting PM10 levels in Kota Bharu, Kelantan. Among the models employed, the RF model yielded the highest accuracy rate, reaching 98.37% for PM10 classification. For a comprehensive overview of related research, refer to Table 1.

3. Materials and methods

In this section, we present the proposed methodology for air quality prediction, comprising the machine learning models and dataset utilized for experimentation. The workflow of the proposed approach is depicted in Fig. 2. Initially, we acquire the dataset containing various air quality features. Since the dataset may contain missing values, data preprocessing becomes imperative to address this issue. To overcome this challenge, we employ the KNN Imputer in this study. Subsequently, various machine learning models, including the MLP model, are applied to the preprocessed dataset. To facilitate model training and evaluation, the dataset is split into training and testing subsets. The classifiers are then employed to categorize air quality as safe or unsafe for human health.

3.1. Proposed IoT framework architecture

The architecture of the proposed solution is organized into five layers, as depicted in Fig. 1.

3.2. Deployment of sensor nodes

3.3. Proposed IoT framework layer-wise architecture

The proposed IoT framework architecture consists of five layers:

1. **Sensor Deployment Layer**: Pollution-detecting sensor nodes, such as PM2.5, CO, and MQ-series sensors, along with a controller and a 5G Network/Wi-Fi module, are deployed in the urban area to measure air pollution levels. A Kalman filter is employed to filter out noise from the sensor data.
2. **Connectivity Layer**: Sensor nodes establish connectivity with the Google cloud server through a Smart Gateway, enabling seamless communication between the nodes and the cloud server.
3. **Data Transmission Layer**: The collected data is transmitted from Firebase in the form of a JavaScript Object Notation (JSON) file. Before further analysis, missing values or erroneous data entries are rectified within the dataset.
4. **Data Analysis Layer**: Various machine learning and ensemble learning models, including the proposed SV-CNN model, are applied to the preprocessed datasets.
5. **Performance Evaluation Layer**: The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F-score. The performance of the SV-CNN model is compared with existing models, as discussed in Section 4.

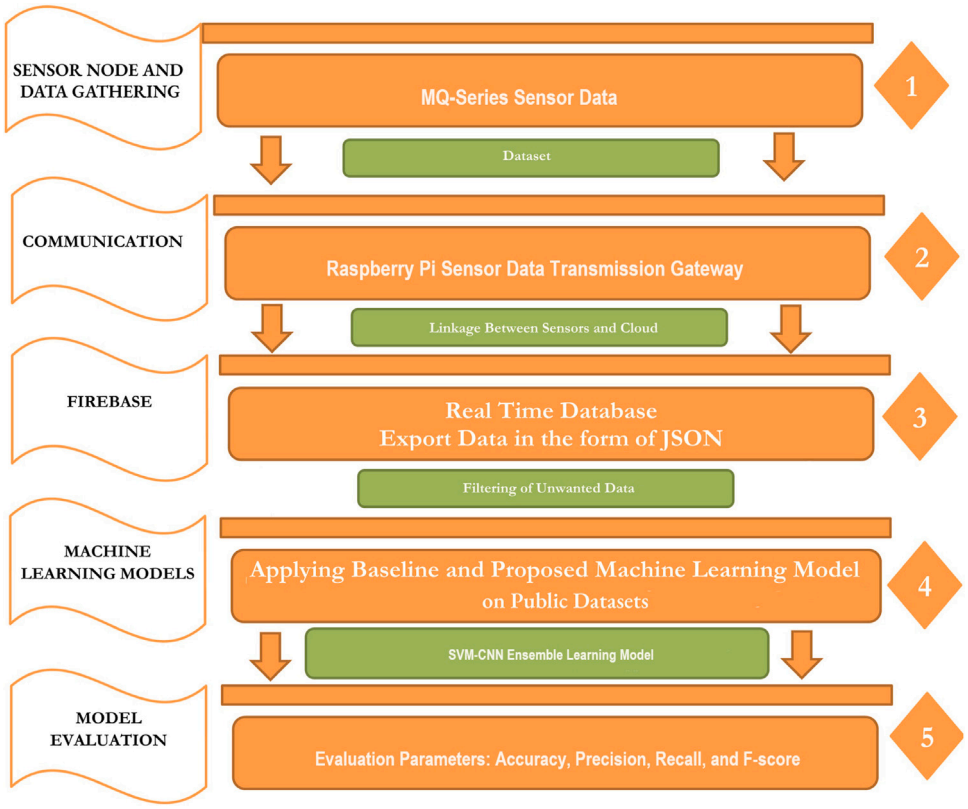

Fig. 1. IoT based smart framework for air quality control.

3.4. Intelligent gateway

Positioned within the urban landscape, the proposed sensor node connects to a Smart Gateway. The Smart Gateway, implemented using Raspberry Pi, establishes a secure and robust connection between the air monitoring device and the cloud server via WiFi or the SIMCom 5G module, compatible with Raspberry Pi [29,30]. Incorporating a 5G network/Wi-Fi module in the deployment of sensor nodes is essential to enable efficient communication and data transmission in modern sensor networks. The SIMCom 5G module is compatible with the Raspberry Pi, providing a seamless integration for establishing connectivity in various applications, including sensor node deployment. These wireless communication technologies offer high-speed data transfer rates and low latency, ensuring real-time transmission of sensor data to the cloud or central server. By leveraging a 5G network/Wi-Fi module, sensor nodes can transmit large volumes of data quickly and reliably, facilitating timely monitoring and analysis of environmental parameters such as air quality, temperature, and humidity. Additionally, the widespread availability and compatibility of 5G networks and Wi-Fi infrastructure make them ideal choices for sensor node deployment in urban and remote areas alike. This integration enhances the scalability, flexibility, and connectivity of sensor networks, ultimately leading to more effective and responsive environmental monitoring systems. The Raspberry Pi serves as the central processing unit, managing data acquisition, preprocessing, and transmission tasks, while the SIMCom 5G module handles the wireless connectivity aspect. Together, they form a robust and versatile platform for building connected devices and systems that leverage the benefits of 5G technology. Primarily, the IoT gateway facilitates communication between nodes and between nodes and the cloud, ensuring the integrity and reliability of the communication link [31–34]. Conceptually, a gateway serves as a hardware device that facilitates data transmission from the source to the destination [35]. Notable features of the Smart Gateway encompass visualization, short-term data retention, network security management, and system diagnostics. Firebase, a platform based on Google web development and mobile technology [36], serves as the backbone for data hosting within the Smart Gateway, with two servers hosting the data: (1) Apache server (webserver) and (2) Node.js server (IoT gateway). The operational workflow of the IoT services is delineated as follows:

1. Deployed sensor nodes detect sensor value variations, treating them as events, and transmit them to the Smart Gateway. The Gateway collects and preprocesses sensor data, applying pre-filtering and refinement procedures. Each sensor reading undergoes comparison with predetermined threshold values and is timestamped before being forwarded to Google Cloud services. Google Cloud processes and stores the received events, with a web server facilitating the generation and transfer of credentials to the web browser. These credentials are then relayed from the browser to the Apache server, functioning as a

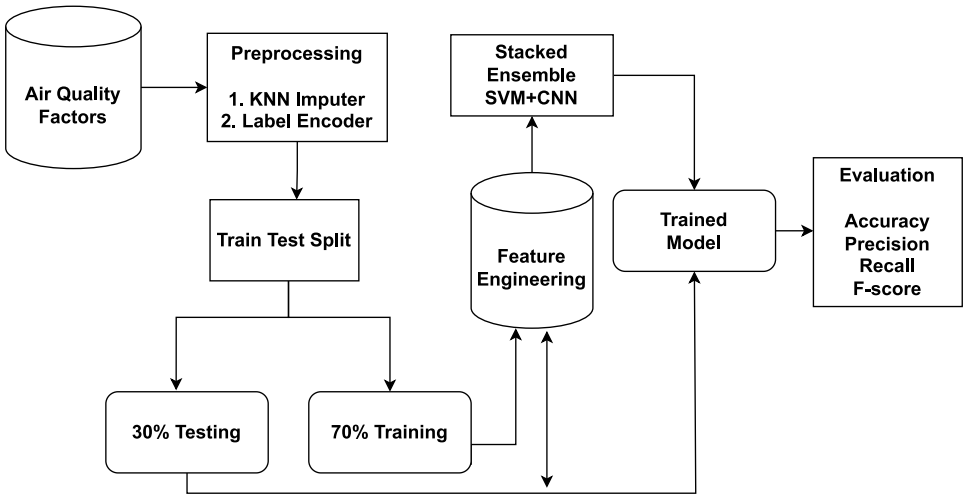

Fig. 2. Workflow diagram of the proposed methodology.

reverse proxy bridge, and further to the Node.js server via HTTP post requests. The HTTP server deployed in the IoT gateway handles authentication of Firebase via the provided credentials. The real-time database, integral to the Firebase platform, operates as a cloud-hosted database capable of storing data in JSON format. This database supports real-time synchronization with client devices across multiple platforms, facilitating both online and offline data access. JSON, an open standard format file, facilitates language-independent data exchange through human-readable text representations of data objects, including data-type arrays and value pairs [37]. The collected data, stored as JSON objects within the Firebase database, can be exported as JSON files. This data serves as a valuable resource for the application of machine learning models to enhance predictive capabilities in urban or smart city environments [38].

The Smart Gateway plays a pivotal role in enabling seamless communication between sensor nodes and the Google Cloud server within the urban landscape. Implemented using Raspberry Pi, the Smart Gateway establishes a secure and robust connection, either via WiFi or the SIMCom 5G module, ensuring reliable data transmission [29,30]. Acting as an IoT gateway, it facilitates communication between nodes and between nodes and the cloud, ensuring the integrity and reliability of the communication link [31–34]. Notable features of the Smart Gateway include visualization, short-term data retention, network security management, and system diagnostics. To ensure confidentiality, integrity, and availability of transmitted data, robust security measures are implemented, including encryption techniques, authentication mechanisms, and access control policies [31–34]. Furthermore, Firebase, serving as the backbone for data hosting within the Smart Gateway, employs two servers — an Apache server and a Node.js server, for data storage and processing. The real-time database in Firebase supports data synchronization across multiple platforms, facilitating both online and offline data access. By integrating sensor data with cloud-based analytics and machine learning algorithms, valuable insights and predictive capabilities are derived, enabling informed decision-making in urban or smart city environments [38]. This integration allows for real-time monitoring, predictive maintenance, and optimization of resources, ultimately enhancing operational efficiency and sustainability.

3.5. Dataset for experiments

Poor air quality in India has emerged as a major health concern and a serious obstacle to economic growth. Dalberg Advisors, a UK-based non-profit management consultancy, produced a recent analysis in partnership with the Industrial Development Corporation, which found that air pollution causes yearly economic losses of up to Rs 7 lakh crore ($95 billion) in India [39]. The main contributors to pollutant emissions in India encompass various sectors such as the energy production industry, vehicular traffic, soil and road dust, waste incineration, power plants, and open waste burning. This study focuses on the analysis of air pollution data sourced from the Central Pollution Control Board (CPCB) in India. The dataset [40] utilized herein comprises 29,531 instances collected from 23 distinct Indian cities spanning from January 2015 to July 2020. Featuring 16 attributes, this dataset furnishes insights into various pollutants across the aforementioned cities. The cities covered include Aizawl, Ahmedabad, Amritsar, Amaravati, Bhopal, Bangalore, Brajrajnagar, Chennai, Chandigarh, Coimbatore, Delhi, Gurugram, Ernakulam, Guwahati, Jaipur, Hyderabad, Kochi, Jorapokhar, Kolkata, Mumbai, Patna, Shillong, Lucknow, Thiruvananthapuram, Talcher, and Visakhapatnam. The attributes encompass Date (YYYY–MM–DD), City, PM10, PM2.5, NO2, NO, NH3, NOx, SO2, CO, Benzene, O3, AQI, Toluene, and AQI_Bucket. This study primarily aims to analyze and forecast the Air Quality Index (AQI) by examining key air pollutants such as PM10, PM2.5, CO, NO2, O3, and SO2. The methodological framework adopted in this study is delineated in the proposed methodology diagram.

Table 2
Missing values details in each attribute.

Attributes	Number of null values	% of null values
Date	0.0‘	0.0
Year	0.0‘	0.0
Month	0.0‘	0.0
NO	3582.0	12.1
NO2	3585.0	12.1
O3	4022.0	13.6
CO	2059.0	7.0
AQI	4681.0	15.9
AQI_Bucket	4681.0	15.9
NOx	4185.0	14.2
PM2.5	4598.0	15.6
Benzene	5623.0	19.0
Toluene	8041.0	27.2
PM10	11 140.0	37.7
NH3	10 328.0	35.0
Xylene	18 109.0	61.3
Particulate Matters	11 899.0	40.3
B_X_o3_NH3	22 788.0	77.2
SO2	3854.0	13.1

3.6. Data preprocessing

Data preprocessing is a crucial step to improve the performance of machine learning models. Redundant or unnecessary data are eliminated in this step as they have no significance for the models. Preprocessing aids in the improvement of learning models' efficiency and reduces computational time. However, missing values are commonly found during data preprocessing. In this research, we discovered several missing values in the dataset during the preprocessing stage. In order to prepare the data for modeling and gain a better understanding of it, a data cleaning process was conducted on the raw data. The first step involved identifying missing values in the dataset. It was discovered that pollutants such as B, X, T, O3, and NH3 had the highest number of missing values, while pollutants like NO, CO, SO2, NO2, NOx, O3, PM 2.5, and AQI had relatively fewer missing values. The missing values were removed using the Pandas 'dropna()' function. To avoid redundancy, the date, city, Year_Month, and AQI_Bucket fields were removed as substitute fields were available in the dataset. The complete missing values details in each attribute are shown in Table 2.

Table 2 indicates that a significant number of values are missing in the dataset. As the dataset is categorical in nature, there are two common methods for handling missing values:

- Fill up the missing values using KNN Imputer.
- Remove the rows containing missing values.

A brief description of both these methods is given in this section of the study.

3.6.1. KNN imputer

In today's data-driven world, information is collected from various sources to gain insights, validate theories, and perform analysis. However, missing data can often occur due to issues with data collection or extraction, resulting in incomplete datasets. Therefore, handling missing values is a critical step in data preprocessing. The choice of imputation method is essential as it can significantly impact the performance of models. One widely used technique for imputing missing values is the KNN Imputer, provided by the scikit learn library. This method is a popular alternative to conventional imputation techniques [41]. The KNN Imputer employs the Euclidean distance matrix to identify the nearest neighbors and impute missing values within the dataset. It computes the Euclidean distance, assigning greater importance to non-missing coordinates while disregarding missing values. The Euclidean distance is determined using the following equation:

$$D_{xy} = \sqrt{weight * squared\ distance\ from\ present\ coordinates} \tag{1}$$

where

$$weight = \frac{total\ number\ of\ coordinates}{number\ of\ present\ coordinates} \tag{2}$$

KNN imputation can perform well for filling in missing values in a dataset under certain conditions. This work considers the following aspects

1. Preservation of Local Patterns: KNN imputation leverages the idea that similar data points should have similar values. By considering the 'k' nearest neighbors, KNN imputation aims to preserve local patterns and relationships in the data. This can be particularly useful when missing values are not missing completely at random but have some underlying structure or dependency on nearby data points.

2. Flexibility: KNN imputation can handle both numerical and categorical data. This flexibility makes it suitable for a wide range of datasets with mixed data types, which is often the case in real-world applications.
3. Adaptability: KNN imputation does not make strong assumptions about the distribution of the data. It adapts to the local characteristics of the dataset, which means it can work well for datasets with complex or nonlinear relationships.
4. Parameter Tuning: The choice of the 'k' value in KNN imputation allows for some degree of control over the imputation process. By selecting an appropriate 'k' value through cross-validation or other techniques, you can fine-tune the imputation to match the characteristics of your dataset.
5. Intuitive Interpretation: KNN imputation is conceptually straightforward to understand. It imputes missing values based on the values of similar data points, making the imputed values interpretable and often aligning well with human intuition.

3.7. Removing missing values from dataset

Another approach for handling missing data is to remove the observations that contain missing values. This means that any data point that has at least one missing value will be removed from the dataset. In the second set of experiments, this approach is used, where all the instances with missing values are removed from the dataset. This can be a straightforward solution to handle missing data, but it can lead to a significant reduction in the size of the dataset and potentially a loss of important information and can affect the performance of the models substantially.

3.8. Machine learning models used for air quality prediction

Machine learning algorithms are crucial in improving the accuracy and effectiveness of air quality classification. There are numerous algorithms available to classify air quality. The Scikit-learn library in Python offers a variety of machine learning classifiers. This open-source library has a large user community and significantly contributes to the research community. In this study, SVC, LR, RF, DT, ETC SGD, KNN, and XGBoost algorithms are implemented using the Scikit-learn library.

Decision Trees (DT) represent a widely utilized machine learning algorithm, often applied to address regression and classification challenges [42]. A critical aspect in DT construction involves selecting the root node at each level, known as "attribute selection". Two prominent techniques for attribute selection are the "Gini index" and "information gain". This study opts for employing the Gini index.

Random Forest (RF) stands as a supervised learning technique, representing an enhanced version of the DT algorithm [43,44]. Compared to alternative classifiers, RF boasts a lower error rate due to its utilization of multiple interconnected DTs and a voting mechanism. To optimize classification, this research employs the Gini index as a cost function for dataset partitioning.

The Logistic Regression (LR) model proves adept at handling a large number of features owing to its straightforward equation for binary classification. LR's hypothesis function calculates the event probability [43,45]. A sigmoid function transforms the LR output into a probability value.

Support Vector Classification (SVC) is a supervised learning algorithm primarily used for pattern-based classification tasks. It operates on the principles of finding the optimal hyperplane that separates different classes of data points in feature space. The key idea behind SVC is to identify the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors [46]. By maximizing the margin, SVC aims to achieve the best possible generalization performance by maximizing the separation between different classes and minimizing the risk of misclassification. In constructing the linear hyperplane, SVC seeks to find the decision boundary that best separates the classes while minimizing classification errors. This decision boundary is defined by a linear equation in the feature space, represented as $w^T x + b = 0$, where w is the weight vector, x is the feature vector, and b is the bias term. The weight vector w determines the orientation of the hyperplane, while the bias term b controls its position in the feature space. To find the optimal hyperplane, SVC solves an optimization problem that involves maximizing the margin subject to the constraint that all data points are correctly classified. This optimization problem can be formulated as a convex quadratic programming (QP) problem, which can be efficiently solved using optimization techniques such as the Sequential Minimal Optimization (SMO) algorithm. In cases where the classes are not linearly separable, SVC can still construct a linear hyperplane in a higher-dimensional space through a process called the kernel trick. By mapping the original feature space into a higher-dimensional space using a nonlinear mapping function (kernel function), SVC can transform the data into a space where classes become separable by a hyperplane. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels, each suited for different types of datasets. Overall, SVC operates by constructing a linear hyperplane in feature space that maximally separates different classes of data points, achieving optimal generalization performance by maximizing the margin between classes. Through the use of kernel functions, SVC can handle both linearly separable and nonlinearly separable datasets, making it a versatile and powerful classification algorithm.

XGBoost emerges as a high-speed supervised learning algorithm utilized in this study for precise air quality classification [47]. It incorporates regularized learning features, facilitating the smoothing of final weights and preventing overfitting.

Extra Tree Classifier (ETC) represents an ensemble learning approach employed for classification tasks [48]. Similar to the RF classifier, ETC constructs multiple decision trees and averages their outputs to enhance model accuracy and robustness. While RF randomly selects feature subsets to determine the best split point at each node, ETC selects split points randomly across the feature space. This characteristic enhances training speed and reduces susceptibility to overfitting.

Stochastic Gradient Descent (SGD) serves as a popular optimization technique iteratively learning optimized parameter values to minimize the cost function (cf) [49]. SGD requires less training time to determine the cost function of a single training sample x_i

at each iteration to reach the local minimum. Multiple hyperparameters are utilized to optimize SGD performance on the analyzed data.

Convolutional Neural Networks (CNN) have exhibited remarkable results in text classification [50]. In the CNN architecture, the initial layer incorporates word2vec embedding for data processing. Classification is executed by the convolutional layer utilizing variable filter sizes, followed by a pooling layer. The softmax layer yields final results. Although CNN excels in text classification, it may encounter challenges in handling long-term word dependencies within sentences, as it primarily focuses on local features.

Algorithm 1 SV-CNN ensembling model.

Input: input data $(x, y)_{i=1}^{N}$
M_{SVM} = Trained_SVM
M_{CNN} = Trained_CNN

1: **for** $i = 1$ *to* M **do**
2: **if** $M_{SVM} \neq 0$ & $M_{CNN} \neq 0$ & $training\_set \neq 0$ **then**
3: $ProbSVC - 1 = M_{SVC}.probability(1 - class)$
4: $ProbSVC - 2 = M_{SVC}.probability(2 - class)$
5: $ProbCNN - 1 = M_{CNN}.probability(1 - class)$
6: $ProbCNN - 2 = M_{CNN}.probability(2 - class)$
7: Decision function = $max(\frac{1}{N_{classifier}} \sum_{classifier}$
$(Avg_{(ProbSVC-1,ProbCNN-1)},$
$(Avg_{(ProbSVC-2,ProbCNN-2)})$
8: **end if**
9: Return final label $\hat{p}$
10: **end for**

3.9. Air quality prediction proposed approach

The dataset used in this study was sourced from Kaggle, a renowned repository recognized for its wide assortment of publicly accessible datasets. A preprocessing phase was performed to solve the problem of missing values and improve the performance of the learning model. KNN Imputer was used to effectively handle the missing values. The dataset was then split in the 70:30 ratio, with 70% used to train the model and 30% for testing. The proposed air quality prediction system employed the SVM+CNN ensemble approach known as SV-CNN. This ensemble model combines one machine learning and one deep learning algorithm, leveraging the strengths of each. Ensemble models are renowned for their capability to amalgamate predictions from multiple models, thereby enhancing accuracy and robustness. Each constituent model within an ensemble contributes its unique strengths, and through their combination, an improved overall performance is achieved.

Lines 3–6 of Algorithm 1 show the probability scores attributed to classes 1 and 2 by the SVC and CNN models, accordingly. The probability score, which is often used in classification problems, represents the likelihood or confidence that a given data sample belongs to a particular class. It is computed using the output of classifier models, such as SVC and CNN in this case. However, this probability does not represent the final prediction for a single target class; rather, it is a raw score that reflects prediction confidence. Certain classifiers include a probability calibration phase that converts raw scores into probabilities. This calibration phase guarantees that the computed scores are correctly calibrated and interpreted as probabilities. Based on the likelihood scores and the predefined threshold (0.5 in this case), the classifier determines a final predicted class label for the data sample. The decision function on line 7 of Algorithm 1 selects the final class with a higher probability score than the set threshold. A practical illustration of how the decision function works is provided below for more clarification.

An example can help to explain how the SV-CNN model works. Each sample processed by both the SVM and CNN models is assigned a probability score. For instance, if the SVM model provides probability scores of 0.5 and 0.6 to classes 1 and 2, respectively, while the CNN model assigns scores of 0.6 and 0.7 to the same classes, the final probabilities are calculated as follows:

$$P(1) = \frac{0.5 + 0.6}{2} = 0.55$$

$$P(2) = \frac{0.6 + 0.7}{2} = 0.65$$

The final class label is determined by the SV-CNN technique, which uses the probability scores for each class from both voting models. Line 7 of Algorithm 1 selects the label with the highest average probability.

Ensemble models work by combining predictions from numerous different learning algorithms. The conventional process for creating an ensemble model entails training multiple models on the same dataset and then combining their predictions. The SV-CNN ensemble model takes this strategy, training SVM and CNN models separately on the same dataset. Each model computes the

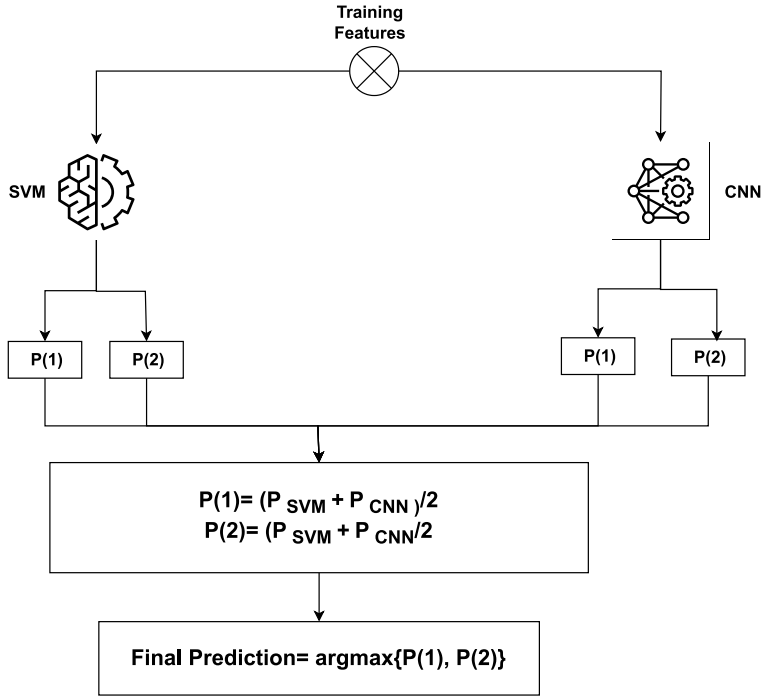

Fig. 3. Architecture of the proposed voting classifier.

estimated probability for each class of the target variable. A final prediction for all the observations in the dataset can then be obtained by combining all these estimated probabilities. Calculating a weighted average of the predicted probabilities – the weights being established by each model's performance on a validation set – is a popular method for merging predictions.

To produce predictions that are more reliable and accurate, the suggested ensemble model makes use of the advantages of both machine learning and deep learning techniques. We improve the model's generalization performance and reduce overfitting by training multiple models on the air quality dataset and combining their predictions. The suggested ensemble model's performance is clarified by Algorithm 1, which can be briefly summarized as follows:

$$\hat{p} = argmax\{\sum_{i}^{n} SVC_i, \sum_{i}^{n} CNN_i\}. \tag{3}$$

where $\sum_{i}^{n} SVC_i$, and $\sum_{i}^{n} CNN_i$, each of the classifiers furnishes prediction probabilities for every test sample. Subsequently, the probabilities assigned to each test instance by the SVC and CNN models undergo evaluation through the soft voting criterion, as illustrated in Fig. 3.

The ensemble model selects the ultimate class by identifying the highest average probability among the classes and amalgamating the predicted probabilities from both classifiers. The final prediction corresponds to the class with the highest probability score, as it is determined by:

$$SV - CNN = argmax(g(x)) \tag{4}$$

3.10. Evaluation metrics

In this study, the efficacy of the proposed system was evaluated through various performance measures, including accuracy, precision, recall, and F1 score. These metrics are frequently utilized in machine learning to assess the accuracy of model predictions. Accuracy quantifies the number of correct predictions relative to the total predictions made, precision represents the proportion of true positive predictions within all positive predictions, recall evaluates the proportion of true positive predictions among all actual positives, and the F1 score combines both precision and recall into a single metric. Mathematically, these metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Table 3
Model results calculated by eliminating missing values.

Model	Accuracy	Precision	Recall	F1 score
LR	73.74	77.55	79.45	78.14
DT	76.41	78.51	80.53	79.76
RF	81.66	80.35	81.56	80.22
SGD	82.94	80.72	81.87	80.65
ETC	82.89	81.52	82.52	81.52
XGB	83.14	81.58	80.53	80.19
SVM	89.52	83.42	83.43	83.42
CNN	86.94	88.43	85.36	86.98
SV-CNN	89.93	90.63	89.45	89.76

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

In the context of these equations, "true positive" (TP) denotes the count of accurately identified positive instances, "false positive" (FP) signifies the count of instances incorrectly labeled as positive, "true negative" (TN) indicates the count of correctly identified negative instances, and "false negatives" (FN) represents the count of instances incorrectly labeled as negative.

3.11. Experimental configuration

For our experiments, we leveraged an Intel Core i7 CPU alongside a NVIDIA graphics processing unit (GPU) for model training. The recommended model was developed within the Python 3.8 programming environment.

4. Results and discussion

This section presents the findings of the air quality classification using various machine learning and ensemble learning classifiers. The performance of these models is assessed based on metrics such as accuracy, precision, recall, and F1 score.

4.1. Results of machine learning models with missing values removed

The first set of experiments involves eliminating missing values from the dataset, followed by the application of machine learning models. Table 3 showcases the outcomes obtained from these models after the removal of missing values from the dataset.

The results indicate that the SVM achieved the highest accuracy of 89.52%, followed closely by the deep learning model CNN with an accuracy of 86.94%. Furthermore, SVM demonstrated a precision of 83.42%, a recall of 83.43%, and an F1 score of 83.42%, whereas CNN exhibited a precision of 88.43%, a recall of 85.36%, and an F1 score of 86.98%. Conversely, LR showed the weakest performance, recording an accuracy of 73.74%, a precision of 77.55%, a recall of 79.45%, and an F1 score of 78.14%. The proposed ensemble model VC (SVM+CNN) surpassed all other individual models, achieving an accuracy of 89.93%, a precision of 90.63%, a recall of 89.45%, and an F1 score of 89.76%. Overall, the performance of individual machine learning models using the dataset with deleted missing values was subpar. The graphical representation in Fig. 4 illustrates the results of machine learning models with deleted missing value data, indicating that except for SVM, CNN, and SV-CNN, the performance of other models is mediocre.

4.2. Performance evaluation of machine learning models employing KNN imputation

The performance is further enhanced through the utilization of the KNN Imputer. During the preprocessing stage, numerous missing values were identified in the dataset. To address this issue, two approaches are typically employed: removing the missing values or imputing them using appropriate techniques. In this study, the KNN Imputer was selected, which utilizes the Euclidean distance formula to identify and fill missing records with appropriate values. The imputed data is then utilized as training data for machine learning models. Another advantage of imputation is the preservation of the total number of records within the dataset, whereas removal of missing values significantly reduces the dataset size. Preserving all values in datasets with limited records is crucial for effective learning. The results obtained using the KNN Imputer in conjunction with machine learning models are presented in Table 4.

The results indicate that SVM and CNN achieved accuracy rates of 97.89% and 96.83%, respectively. These findings support the hypothesis that employing the KNN Imputer to fill missing values enhances model performance, as evidenced by the improved performance of machine learning models compared to using datasets with missing values. The SV-CNN ensemble model proposed in this study achieved the highest accuracy rate of 99.17% among all models. Furthermore, the ensemble model demonstrated a precision of 99.75%, a recall of 99.24%, and an F1 score of 99.52%. In contrast, the linear regression model LR exhibited the lowest accuracy value of 85.75%. Fig. 5 provides a graphical representation of the outcomes of machine learning models using the KNN Imputer, illustrating that employing the KNN Imputer enhances the performance of machine learning models.

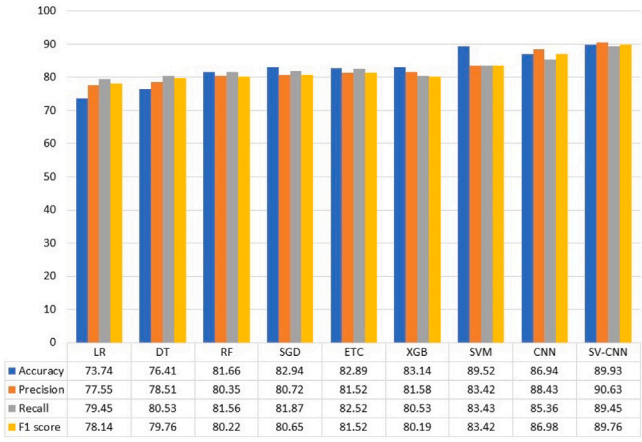

Fig. 4. A summary of the results of machine learning models achieved after removing missing values from the dataset.

Table 4
KNN imputed dataset results.

Model	Accuracy	Precision	Recall	F1 score
LR	85.75	87.34	89.67	88.34
DT	86.45	88.84	90.49	89.29
RF	91.37	90.29	91.22	90.63
SGD	92.59	90.08	91.73	90.45
ETC	92.63	91.46	92.25	91.58
XGB	93.74	91.37	90.30	90.79
SVM	97.89	93.73	93.66	93.14
CNN	96.83	98.88	95.59	96.79
SV-CNN	**99.17**	**99.75**	**99.24**	**99.52**

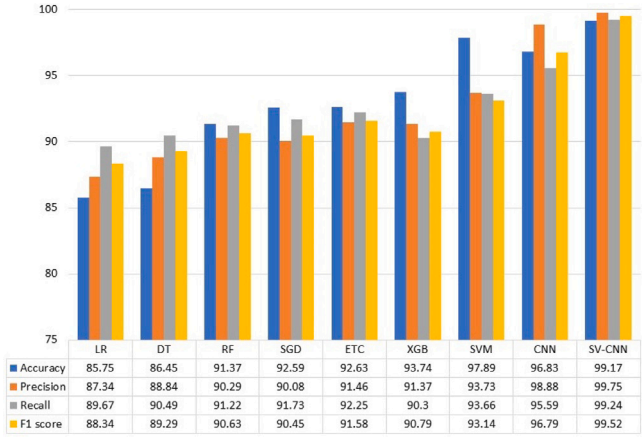

Fig. 5. Results of the learning models using KNN Imputer.

4.3. Assessing the impact of KNN imputer on machine learning models

We compared the performance of machine learning models with and without the KNN Imputer in order to evaluate its efficacy. When the KNN Imputer was used in the second experiment instead of data with removed missing values, the results showed a significant improvement in the performance of machine learning models. Table 5 illustrates the results of the machine learning models for both scenarios, enabling a comprehensive analysis of their performance.

The comparative performance between the two scenarios, where missing values are removed and when using the KNN Imputer, is depicted in Fig. 6. The graph shows that using the KNN Imputer improves the performance of each of the models, resulting in a better overall performance for all machine learning models.

Table 5
Comparative Analysis of Machine Learning Models Using KNN-Imputed Dataset Results.

Model	With KNN	Without KNN
LR	85.75	73.74
DT	86.45	76.41
RF	91.37	81.66
SGD	92.59	82.94
ETC	92.63	82.89
XGB	93.74	83.14
SVM	97.89	89.52
CNN	96.83	86.94
SV-CNN	**99.17**	**89.93**

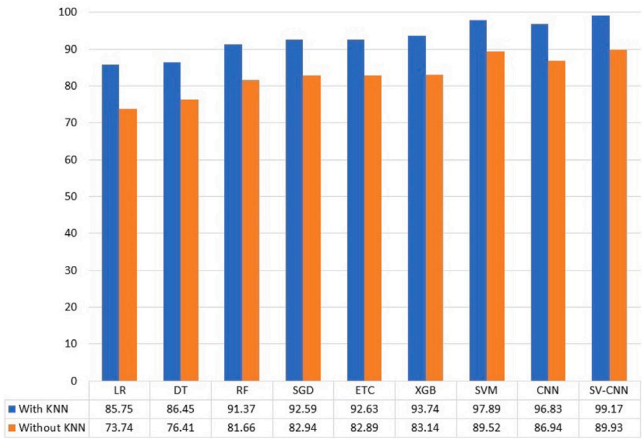

Fig. 6. Machine learning models and KNN imputed dataset results performance comparison.

4.4. K-fold cross-validation results

To ascertain the robustness of the models, K-fold cross-validation was utilized. The outcomes of 5-fold cross-validation are depicted in Fig. 7, showcasing that the proposed approach surpasses other models regarding accuracy, precision, recall, and F1 score, with minimal standard deviation.

4.5. Time complexity of models

Time complexity in machine learning primarily concerns the training phase, indicating the duration needed to adjust the model's parameters based on the input data. This complexity varies depending on factors such as the model's architecture, the size of the training dataset, and the optimization algorithm employed. Table 6 presents the training time of all machine learning models with and without the KNN imputation technique. The table reveals that the ensemble of learning models does not entail significant time consumption, as the voting classifier benefits from the 'n_job' parameter set to −1, utilizing all available cores of the system during training. The computational time for the proposed model is 95 s with KNN, which surpasses that of individual models such as LR, SGD, SVM, and CNN. Nonetheless, the proposed model's accuracy significantly outperforms individual models despite the slightly longer training time.

4.6. Why air quality monitoring is important?

Monitoring air quality is vital due to the potential adverse effects of unregulated pollution on human health and the environment. Poor air quality can lead to respiratory and cardiovascular diseases, exacerbate existing health conditions, and even cause premature death. Additionally, pollutants such as particulate matter (PM), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and ozone (O_3) can have detrimental impacts on ecosystems, including harm to vegetation, wildlife, and aquatic systems. Monitoring programs play a crucial role in identifying patterns and trends in air quality over time and across different geographical areas. By collecting data on pollutant concentrations and meteorological conditions, monitoring programs can assess the effectiveness of air quality regulations and interventions, track changes in pollution levels, and identify emerging environmental and health risks. This information enables policymakers, public health officials, and stakeholders to develop targeted strategies for pollution control and mitigation, ultimately safeguarding human health and the environment [51].

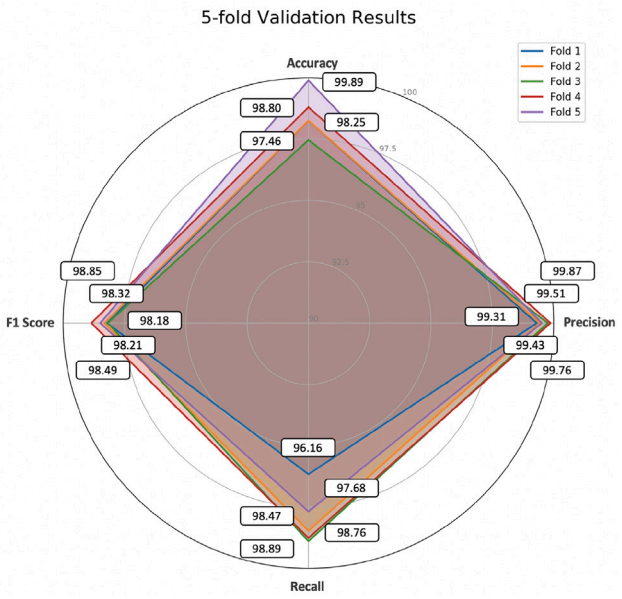

Fig. 7. Proposed approach 5-fold validation results. **Average:** 98.61 & 99.72 & 97.54 & 98.41, respectively for **Accuracy**, **Precision**, **Recall**, and **F1 score**.

Table 6
Learning models time computational complexity (in seconds).

Model	With KNN	Without KNN
LR	87s	80s
DT	96s	89s
RF	102s	94s
SGD	90s	88s
ETC	105s	98s
XGB	115s	103s
SVM	88s	82s
CNN	94s	90s
SV-CNN	95s	89s

4.7. Significance of the proposed model

To assess the significance and stability of the proposed model, an additional dataset sourced from the UCI Machine Learning Repository focusing on PM2.5 concentrations in Beijing is utilized [52]. Spanning from 2010 to 2014, this dataset comprises 43,824 samples, each representing a one-hour duration. It includes parameters such as years, months, dates, hours, PM2.5 concentration levels, dew points, temperatures, pressures, combined wind directions, cumulated wind speeds, cumulated hours of snow, and cumulated hours of rain. Preprocessing techniques are employed to ensure data integrity by identifying and eliminating redundant values. The application of the proposed SV-CNN model to this dataset, coupled with KNN imputation, yields impressive results, with an accuracy score of 99.38%, precision of 98.54%, recall of 98.78%, and F-score of 98.66%. This consistent and reliable performance on a distinct dataset underscores the stability and effectiveness of the proposed model.

4.8. Performance comparison with existing studies

To demonstrate the superiority of the proposed model over previous state-of-the-art approaches, a comparison is conducted with relevant existing studies. Specifically, three studies renowned for employing state-of-the-art models to enhance accuracy are selected for comparison. For instance, [53] utilized SMOTE features with the Random Forest (RF) machine learning model for cervical cancer detection, achieving an accuracy score of 96.06%. Another study [54] employed DBSCAN with SMOTETomek and RF as a machine learning model, achieving the highest accuracy score of 97.72%. Similarly, [55] utilized Recursive Feature Elimination (RFE) with SMOTETomek, reporting an accuracy score of 98.81%. Table 7 compares the performance of the proposed and current research, demonstrating the superiority of the suggested approach.

5. Conclusions

Providing accurate and customized air quality forecasts to urban stakeholders is crucial, especially in areas inundated with traffic, where elevated air pollution levels pose significant health risks. Air quality is paramount for human well-being, particularly

Table 7
Performance comparison with previous studies.

Ref	Classifiers	Achieved accuracy
[24]	NNs	95%
[26]	NNs and SVM	97.3%
[56]	KNN, GNB, SVM,RF, XGBoost	90%
[25]	SVR, RFR, CR	97.6080% RFR
[27]	SMOTEDNN, XGBoost, RF, SVM, KNN	90.9% SMOTEDNN, (careful hyperparameter tuning, which can be a time-consuming process)
Proposed	**SV-CNN**	**99.17% with KNN Imputer**

in urban locales characterized by high vehicular and industrial activities. Clean air is indispensable, and surpassing permissible gas concentrations can jeopardize public health. Thus, monitoring the air quality index and gas concentrations is imperative, ensuring a conducive living environment. This study introduces an automated approach for predicting the air quality index using a machine learning ensemble model. The initial step involves normalizing the dataset using the KNN Imputer technique. Subsequently, the stacked ensemble SV-CNN model is employed. The results, boasting a remarkable accuracy of 99.17%, underscore the effectiveness of ensemble models in air quality prediction. Furthermore, an analysis of other state-of-the-art models highlights the superiority of the proposed approach. Future endeavors aim to further enhance model performance by employing stacked ensembles of machine and deep learning models, facilitating robust and generalized results, particularly on higher-dimensional datasets.

CRediT authorship contribution statement

Khaled Alnowaiser: Writing – review & editings, Conceptualization. **Aisha Ahmed Alarfaj:** Writing – original draft, Visualization, Software. **Ebtisam Abdullah Alabdulqader:** Data curation, Methodology. **Muhammad Umer:** Methodology, Writing – original draft, Conceptualization. **Lucia Cascone:** Writing – review & editing, Resources. **Bhavya Alankar:** Final manuscript review, Project supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R348), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. 2018, URL https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action. [Accessed on 12 April 2023].
[2] Sciencing. Pollution's impact on historical monuments. 2019, URL https://sciencing.com/about-6372037-pollution-s-impact-historical-monuments.html. [Accessed on 12 April 2023].
[3] Fahad Shah, Sonmez Osman, Saud Shah, Wang Depeng, Wu Chao, Adnan Muhammad, et al. Plant growth regulators for climate-smart agriculture. CRC Press; 2021.
[4] Fahad Shah, Sonmez Osman, Saud Shah, Wang Depeng, Wu Chao, Adnan Muhammad, et al. Sustainable soil and land management and climate change. CRC Press; 2021.
[5] Wahid Fazli, Sharif Muhmmad, Ali Amjad, Fahad Shah, Adnan Muhammad, Noor Muhammad, et al. Plant-microbes interactions and functions in changing climate. Environ Clim Plant Veget Growth 2020;397–419.
[6] Wu Jianxin, Feng Ziwei, Ma Chunbo. Promotion incentives and environmental regulation: Evidence from China's environmental one-vote veto evaluation regime. Environ Resour Econ 2024;87(1):257–86.
[7] Riley Rosie, de Preux Laure, Capella Peter, Mejia Cristian, Kajikawa Yuya, de Nazelle Audrey. How do we effectively communicate air pollution to change public attitudes and behaviours? A review. Sustain Sci 2021;1–21.
[8] Shanmugam GS, et al. Smart green resource conservation approach for smart IoT cloud. J Comput Theor Nanosci 2018;15(6–7):2069–75. http://dx.doi.org/10.1166/jctn.2018.7409.
[9] Zhu L, et al. Do economic activities cause air pollution? Evidence from China's major cities. Sustainable Cities Soc 2019;49:101593. http://dx.doi.org/10.1016/j.scs.2019.101593.
[10] Sharma R, et al. Inferring air pollution from air quality index by different geographical areas: Case study in India. Air Quality Atmosphere Health 2019;12(11):1347–57. http://dx.doi.org/10.1007/s11869-019-00749-x.
[11] Zhang H, et al. Air pollution and control action in Beijing. J Clean Prod 2016;112:1519–27. http://dx.doi.org/10.1016/j.jclepro.2015.04.092.

[12] Wang Shuxiao, Zhao Meng, Xing Jia, Wu Ye, Zhou Yu, Lei Yu, et al. Quantifying the air pollutants emission reduction during the 2008 Olympic games in Beijing. Environ Sci Technol 2010;44(7):2490–6.

[13] Monika Poskart, Szecówka Lech. Ecological effect of air/fuel staging and flue gas recirculation on NOX formation–experimental and numerical analysis. J Min Inst 2007;170(2):250.

[14] Bank The World. Urban population (% of total population). 2018, URL https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS. [Accessed on 12 April 2023].

[15] United Nations. Department of economic and social affairs: Urban population change. 2018, https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. [Accessed on 12 April 2023].

[16] Wang Shuxiao, Hao Jiming. Air quality management in China: Issues, challenges, and options. J Environ Sci 2012;24(1):2–13.

[17] Li Xizhe, Zou Nianyu, Wang Zhisheng. Application of a deep learning fusion model in fine particulate matter concentration prediction. Atmosphere 2023;14(5):816.

[18] Mu Bin, Jiang Xin, Yuan Shijin, Cui Yuehan, Qin Bo. NAO seasonal forecast using a multivariate air–sea coupled deep learning model combined with causal discovery. Atmosphere 2023;14(5):792.

[19] Strizhenok Alexey V, Ivanov Andrey V. Monitoring of air pollution in the area affected by the storage of primary oil refining waste. J Ecol Eng 2021;22(1):60–7.

[20] Chebyshev IS, Baryshnikov ES, Legkokonets VA. Application of machine learning to predict the acoustic properties of rock samples. PRONEFT'. Professional'no o nefti 2018;4(10):67–70.

[21] Nahar Khalid MO, Ottom Mohammad Ashraf, Alshibli Fayha, Shquier Mohammed M Abu. Air quality index using machine learning–A Jordan case study. Compusoft 2020;9(9):3831–40.

[22] Castelli Mauro, Clemente Fabiana Martins, Popovič Aleš, Silva Sara, Vanneschi Leonardo. A machine learning approach to predict air quality in California. Complexity 2020;2020.

[23] Sanjeev Dyuthi. Implementation of machine learning algorithms for analysis and prediction of air quality. Int J Eng Res Technol 2021;0181–2278.

[24] Soundari A Gnana, Jeslin J Gnana, Akshaya AC. Indian air quality prediction and analysis using machine learning. Int J Appl Eng Res 2019;14(11):181–6.

[25] Hamami Faqih, Dahlan Iqbal Ahmad. Air quality classification in urban environment using machine learning approach. IOP Conf Ser: Earth Environ Sci 2022;986(1):012004.

[26] Mahalingam Usha, Elangovan Kirthiga, Dobhal Himanshu, Valliappa Chocko, Shrestha Sindhu, Kedam Giriprasad. A machine learning model for air quality prediction for smart cities. In: 2019 international conference on wireless communications signal processing and networking. IEEE; 2019, p. 452–7.

[27] Haq Mohd Anul, et al. Smotednn: A novel model for air pollution forecasting and AQI classification. Comput Mater Contin 2022;71(1):1403–25.

[28] Shaziayani Wan Nur, Ul-Saufie Ahmad Zia, Mutalib Sofianita, Mohamad Noor Norazian, Zainordin Nazatul Syadia. Classification prediction of PM10 concentration using a tree-based machine learning approach. Atmosphere 2022;13(4):538.

[29] You I, et al. Misbehavior detection of embedded IoT devices in medical cyber physical systems. In: Proceedings of the 2018 IEEE/ACM international conference on connected health: Applications, systems and engineering technologies. 2018, p. 88–93.

[30] Sharma V, et al. Behavior and vulnerability assessment of drones-enabled industrial internet of things (IIoT). IEEE Access 2018;6:43368–83. http://dx.doi.org/10.1109/access.2018.2856368.

[31] Sharma V, et al. Security of 5G-V2X: Technologies, standardization, and research directions. IEEE Netw 2020;34(5):306–14. http://dx.doi.org/10.1109/mnet.001.1900662.

[32] Sharma V, et al. Security, privacy and trust for smart mobile-Internet of Things (M-IoT): A survey. IEEE Access 2020;8:167123–63. http://dx.doi.org/10.1109/access.2020.3022661.

[33] Sharma V, et al. BRIoT: Behavior rule specification-based misbehavior detection for IoT-embedded cyber-physical systems. IEEE Access 2019;7:118556–80. http://dx.doi.org/10.1109/access.2019.2917135.

[34] Khan R, et al. A machine learning based energy-efficient non-orthogonal multiple access scheme. In: 14th international forum on strategic technology. 2019, p. 330–5.

[35] Sharma V, et al. MIH-SPFP: MIH-based secure cross-layer handover protocol for fast proxy mobile IPv6-IoT networks. J Netw Comput Appl 2019;125:67–81. http://dx.doi.org/10.1016/j.jnca.2018.09.002.

[36] Firebase, . 2022, https://firebase.google.com/. [Accessed on 11 February 2022].

[37] Crockford Douglas. The application/Json media type for JavaScript object notation (Json). 2006, RFC 4627.

[38] Rani S, et al. Amalgamation of advanced technologies for sustainable development of smart city environment: A review. IEEE Access 2021;9:150060–87. http://dx.doi.org/10.1109/access.2021.3125527.

[39] (DTE) Swagata Dey. Air pollution in India has caused losses of up to Rs 7 lakh crore annually. 2021, URL https://www.downtoearth.org.in/blog/air/air-pollution-in-india-has-caused-losses-of-up-to-rs-7-lakh-crore-annually-76616#:~:text=Published%3A%20Thursday%2022%20April%202021,of%20Indian%20Industry%20(CII). [Accessed on 12 April 2023].

[40] VOPANI Kaggle. Air quality data in India (2015 - 2020). 2020, URL https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india. [Accessed on 12 April 2023].

[41] Juna Afaq, Umer Muhammad, Sadiq Saima, Karamti Hanen, Eshmawi Ala'Abdulmajid, Mohamed Abdullah, et al. Water quality prediction using KNN imputer and multilayer perceptron. Water 2022;14(17):2592.

[42] Manzoor Mubariz, Umer Muhammad, Sadiq Saima, Ishaq Abid, Ullah Saleem, Madni Hamza Ahmad, et al. RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. IEEE Access 2021;9:128359–71.

[43] Breiman Leo. Bagging predictors. Mach Learn 1996;24(2):123–40.

[44] Biau Gérard, Scornet Erwan. A random forest guided tour. Test 2016;25(2):197–227.

[45] Besharati Elham, Naderan Marjan, Namjoo Ehsan. LR-HIDS: Logistic regression host-based intrusion detection system for cloud environments. J Ambient Intell Humaniz Comput 2019;10(9):3669–92.

[46] Sarwat Samina, Ullah Naeem, Sadiq Saima, Saleem Robina, Umer Muhammad, Eshmawi Ala'Abdulmajid, et al. Predicting students' academic performance with conditional generative adversarial network and deep SVM. Sensors 2022;22(13):4834.

[47] Ashraf Imran, Narra Manideep, Umer Muhammad, Majeed Rizwan, Sadiq Saima, Javaid Fawad, et al. A deep learning-based smart framework for cyber-physical and satellite system security threats detection. Electronics 2022;11(4):667.

[48] Umer Muhammad, Sadiq Saima, Nappi Michele, Sana Muhammad Usman, Ashraf Imran, et al. ETCNN: Extra tree and convolutional neural network-based ensemble model for COVID-19 tweets sentiment classification. Pattern Recognit Lett 2022;164:224–31.

[49] Umer Muhammad, Sadiq Saima, Missen Malik Muhammad Saad, Hameed Zahid, Aslam Zahid, Siddique Muhammad Abubakar, et al. Scientific papers citation analysis using textual features and SMOTE resampling techniques. Pattern Recognit Lett 2021;150:250–7.

[50] Hameed Ahmad, Umer Muhammad, Hafeez Umair, Mustafa Hassan, Sohaib Ahmed, Siddique Muhammad Abubakar, et al. Skin lesion classification in dermoscopic images using stacked convolutional neural network. J Ambient Intell Humaniz Comput 2021;1–15.

[51] Pope C Arden, Lefler Jacob S, Ezzati Majid, Higbee Joshua D, Marshall Julian D, Kim Sun-Young, et al. Mortality risk and fine particulate air pollution in a large, representative cohort of US adults. Environ Health Perspect 2019;127(7):077007.

[52] UCI Machine Learning Repository: Beijing PM2.5 Data Data Set. 2022, https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data. [Accessed on 11 February 2022].

[53] Abdoh Sherif F, Rizka Mohamed Abo, Maghraby Fahima A. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access 2018;6:59475–85.
[54] Ijaz Muhammad Fazal, Attique Muhammad, Son Youngdoo. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors 2020;20(10):2809.
[55] Tanimu Jesse Jeremiah, Hamada Mohamed, Hassan Mohammed, Kakudi Habeebah, Abiodun John Oladunjoye. A machine learning method for classification of cervical cancer. Electronics 2022;11(3):463.
[56] Kumar K, Pande BP. Air pollution prediction with machine learning: A case study of Indian cities. Int J Environ Sci Technol 2022;1–16.