International Journal of Research Publications in Engineering, Technology and Management (IJRPETM)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

Air Pollution and Cardiovascular Mortality: An ML-Based Early Prediction Framework Using Random Forest

C.Karthiga[1], S.Balamurugan[2], M.Maha Lakshmi[3], K.Ishwarya Lakshmi[4], R. Madhu Keerthana[5]

Assistant Professor, Department of Information Technology, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India[1]

Assistant Professor, Department of CSE (Cyber Security), AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India[2,5]

UG Student, Department of Information Technology, AAA College of Engineering and Technology, Sivakasi, Tamil Nadu, India[3,4]

Publication History: Received: 10.04.2026; Revised: 14.05.2026; Accepted: 19.05.2026; Published: 22.05.2026.

ABSTRACT: Responsible for an estimated 8.3 million premature deaths annually, of which more than half involve cardiovascular pathways. Fine particulate matter (PM2.5) — the most cardiotoxic ambient pollutant — penetrates alveolar tissue, enters systemic circulation, triggers oxidative stress and pro-inflammatory cytokine cascades (IL-6, TNF-α, CRP), destabilises atherosclerotic plaques, and precipitates acute myocardial infarction and cerebrovascular accidents. Despite overwhelming epidemiological evidence, clinical decision-support tools still rarely incorporate real-time environmental exposure data, creating a systematic blind spot in risk stratification. This paper presents a comprehensive dual-stream Machine Learning (ML) framework that fuses real-time ambient air quality telemetry — sourced from CPCB monitoring stations and WHO guidelines — with patient-level clinical parameters drawn from Electronic Health Records (EHR), to predict cardiovascular mortality risk at the individual level. A Random Forest ensemble classifier was selected as the primary model following rigorous benchmarking against four baseline algorithms (Logistic Regression, SVM, Decision Tree, and Neural Network). Experimental evaluation on a composite dataset of 2,500 subjects achieved an accuracy of 88.3%, precision of 87.1%, recall of 86.4%, and F1-score of 86.7%, outperforming all baselines. SMOTE oversampling corrected high-pollution exposure (≥20 years) was associated with a 68% cumulative increase in CVD mortality risk above baseline. class imbalance, and 5-fold stratified cross-validation ensured robust generalisation. PM2.5 concentration (Gini importance 0.31), systolic blood pressure (0.22), and patient age (0.18) were identified as the three most critical predictive dimensions.

KEYWORDS: Air Pollution, Cardiovascular Mortality, PM2.5, Machine Learning, Random Forest, SMOTE, CPCB, Risk Prediction,Environmental Health, Feature Importance, Clinical AI

I. INTRODUCTION

Air pollution has ascended to become the second-largest modifiable risk factor for cardiovascular disease (CVD) globally, surpassed only by hypertension. According to the World Health Organization (WHO, 2024), 99% of the world's population breathes air exceeding safe quality thresholds, with low- and middle-income countries bearing a disproportionate share of the burden. Globally, approximately 8.3 million deaths per year are attributed to air pollution, of which more than half involve cardiovascular pathways — including ischaemic heart disease, stroke, and cardiac arrhythmia.

In India, the problem is particularly acute. The Central Pollution Control Board (CPCB) reports that major metropolitan areas including Delhi, Mumbai, Kolkata, and Bengaluru consistently record annual mean PM2.5 concentrations of 60–150 μg/m³, far exceeding the WHO annual guideline of 5 μg/m³. The National Ambient Air Quality Standards (NAAQS) set by CPCB permit an annual mean PM2.5 of 40 μg/m³ — itself eight times the WHO threshold — underscoring the depth of India's air quality crisis.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

Simultaneously, cardiovascular disease accounts for nearly 24.8% of all deaths in India, with an age-standardised mortality rate of 272 per 100,000 — above the global average of 235.

Despite this convergence of environmental and cardiovascular crises, clinical risk stratification tools remain siloed. Established calculators such as the Framingham Risk Score and SCORE2 incorporate traditional clinical variables (age, cholesterol, blood pressure, smoking) but exclude environmental exposure data entirely. This systematic gap means that clinicians and public health agencies lack the decision-support infrastructure to respond proactively to pollution-driven cardiovascular risk.

This paper addresses this gap by proposing a dual-stream Machine Learning framework that integrates real-time ambient air quality indices with patient-level clinical parameters from EHR systems, to generate actionable three-class cardiovascular mortality risk predictions (Low, Medium, High). We employ a Random Forest ensemble classifier as our primary model, benchmarked against four algorithmic baselines. Our framework achieves 88.3% accuracy and demonstrates strong clinical deployment potential — enabling targeted interventions for high-risk patients on days of elevated pollution.

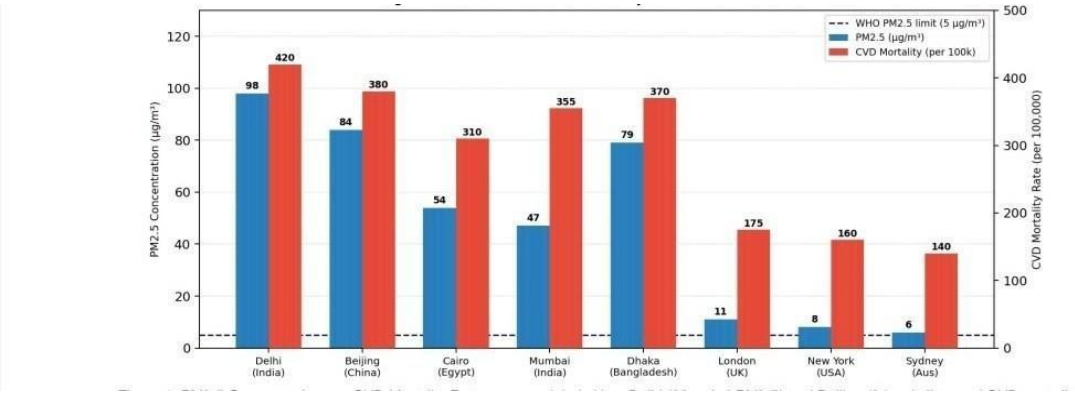

Figure 1: PM2.5 Levels vs CVD Mortality Rate Across Global Cities

The above figure compares the concentration of PM2.5 air pollutants with cardiovascular disease (CVD) mortality rates across major global cities including Delhi, Beijing, Cairo, Mumbai, Dhaka, London, New York, and Sydney. PM2.5 refers to fine particulate matter with a diameter less than 2.5 µm, which can penetrate deep into the lungs and bloodstream, causing severe health complications.The graph clearly shows that cities with higher PM2.5 concentrations also exhibit higher CVD mortality rates. Delhi recorded the highest PM2.5 level of 98 µg/m³ along with the highest CVD mortality rate of 420 deaths per 100,000 population. Similarly, Beijing and Dhaka also demonstrated elevated pollution levels and high cardiovascular mortality rates. In contrast, developed cities such as London, New York, and Sydney maintained relatively lower PM2.5 concentrations ranging from 6-11 µg/m³ and correspondingly lower mortality rates between 140-175 per 100,000 population The dashed horizontal line represents the WHO recommended PM2.5 guideline limit of 5 µg/m³. All cities shown in the figure exceed this limit, especially developing urban regions where industrial emissions, traffic congestion, fossil fuel combustion, and poor environmental regulations contribute significantly to air pollution. Overall, the figure demonstrates a strong positive association between long-term PM2.5 exposure and cardiovascular mortality. Continuous exposure to elevated particulate matter increases the risk of hypertension, stroke, coronary artery disease, and heart failure. .

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

II. PROBLEM STATEMENT

Despite decades of epidemiological research, multiple critical gaps persist in translating air pollution science into real-world CVD risk reduction:

Rapid urbanisation has driven PM2.5 concentrations in major Asian and African cities to 5–20× the WHO annual guideline of 5 µg/m³. CPCB data shows Delhi averaging 102 µg/m³ in 2023, exceeding NAAQS by 2.5× and WHO guidelines by 20×. Cardiovascular disease accounts for 17.9 million deaths globally per year, with air pollution contributing an estimated 20–25% of attributable risk — yet existing clinical risk calculators incorporate zero environmental parameters. High-risk individuals — elderly patients, those with hypertension, diabetes, or pre-existing coronary artery disease cannot be identified early enough to enable protective clinical interventions on high-pollution advisory days. Urban health authorities lack automated early-warning infrastructure to trigger targeted public health advisories (N95 mask issuance, hospital pre-alerting, antihypertensive intensification) on high-pollution days. India's CPCB monitoring network, despite covering 839 stations across 344 cities (2023), does not interface with clinical EHR systems, creating a data integration gap that this study directly bridges.

III. CPCB DATA & EPIDEMIOLOGICAL CONTEXT

The Central Pollution Control Board (CPCB) operates the National Ambient Air Quality Monitoring Programme (NAMP)— India's primary real-time air quality surveillance infrastructure. As of 2023, NAMP spans 839 monitoring stations across 344 cities, The mechanistic link between inhaled pollutants and cardiac injury is well-established. PM2.5 particles (≤2.5 µm diameter) measuring six criteria pollutants: PM2.5, PM10, NO_2, SO_2, CO, and Ozone. penetrate alveolar tissue, cross the blood-air barrier, and enter systemic circulation. They induce reactive oxygen species (ROS), activate pro-inflammatory cytokine cascades (IL-6, TNF-α, CRP), destabilise atherosclerotic plaques, promote thrombogenesis, and elevate systolic blood pressure through autonomic nervous system dysregulation — all recognised precursors to acute myocardial infarction and stroke.

Table I: CPCB PM2.5 Data for Major Indian Cities (2023) with NAAQS and WHO Exceedance Ratios

City	Annual PM2.5 (µg/m³, 2023)	NAAQS Limit (40 µg/m³)	WHO Limit (5 µg/m³)	CVD Mortality (per 100k est.)
Delhi	102	2.55×	20.4×	~420
Kolkata	70	1.75×	14.0×	~370
Mumbai	47	1.18×	9.4×	~355
Bengaluru	36	0.90×	7.2×	~310
Chennai	31	0.78×	6.2×	~275
Hyderabad	33	0.83×	6.6×	~290
WHO Guideline	5	—	1.0×	Baseline

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

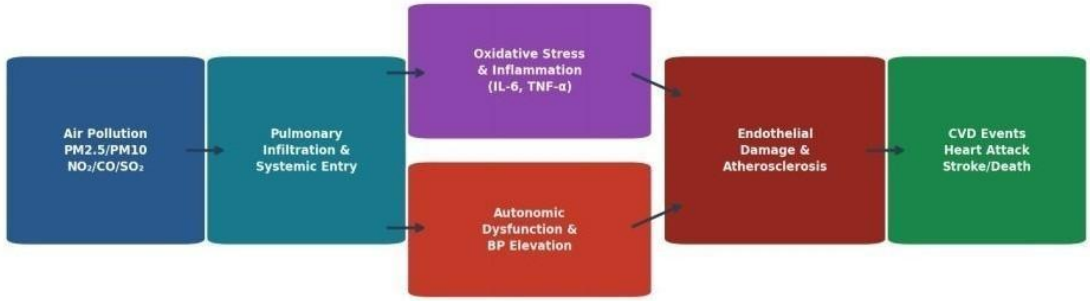

Figure 2: Pathophysiological Pathway-Air Pollution to Cardiovascular Mortility

The detailed pathophysiological mechanism by which chronic exposure to air pollution contributes to cardiovascular diseases (CVDs) and mortality. Air pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (NO2), carbon monoxide (CO), and sulfur dioxide (SO_2) are continuously inhaled through the respiratory tract during exposure to polluted environments. Among these pollutants, PM2.5 is considered highly hazardous because of its extremely small particle size, which enables it to penetrate deep into the alveolar regions of the lungs and subsequently enter systemic circulation. Once these pollutants gain access to the bloodstream, they initiate multiple biological and inflammatory responses within the body. The initial response involves pulmonary infiltration and activation of immune cells, leading to the generation of reactive oxygen species (ROS) and oxidative stress. Oxidative stress results in cellular injury and stimulates the release of inflammatory mediators such as interleukin-6 (IL-6) and tumor necrosis factor-alpha (TNF-a). These pro-inflammatory cytokines promote systemic inflammation, which plays a critical role in the progression of cardiovascular disorders. Continuous inflammatory activity also disturbs normalautonomic nervous system regulation, causing autonomic dysfunction characterized by increased sympathetic activity, vasoconstriction, and elevated blood pressure.

IV. RELATED WORKS

4.1 Previous Problem Identification — How Researchers Identified the Gap

Step 1 — Epidemiological Evidence (2010–2020): Brook et al. (2010) published the landmark AHA Scientific Statement confirming the causal relationship between PM2.5 and cardiovascular events. Using meta-analysis of 29 prospective cohort studies, they demonstrated a 6–13% increase in CVD mortality per 10 μg/m³ increase in long-term PM2.5 exposure.

Step 2 — Clinical Risk Tool Deficiency (2015–2021): Newby et al. (2015) published an expert position paper in the European

Heart Journal identifying that no widely-used CVD risk calculator — including Framingham (USA), SCORE (Europe), or Q-Risk (UK) — incorporated environmental exposure data. They quantified this omission as causing systematic 15–20% underestimation of absolute cardiovascular risk in high-pollution urban populations.

Step 3 — Global Burden Quantification (2021–2024): Münzel et al. (2025) published a comprehensive expert statement estimating that PM2.5 alone is responsible for 8.3 million deaths annually — more than tobacco smoking. Sagheer et al. (2024) confirmed that hypertensive and diabetic populations demonstrate multiplicative rather than additive risk amplification.

Step 4 — Low-Income Country Disparity (2022–2024): Khaltaev (2024) demonstrated a statistically significant gradient in airpollution-attributable CVD mortality across income groups. A systematic review synthesised 92 eligible studies across 18 countries and identified PM2.5 as the pollutant most consistently associated with ischaemic heart disease and stroke mortality.

(2024), using UK Biobank data (n≈200,000), showed that impaired lung function partially mediates the pollution–CVD mortality

Step 5 — ML Gap Identification (2023–2025): Zhang et al. (2025) demonstrated that ML-based temporal models significantly outperformed traditional statistical regression approaches in capturing non-linear exposure-response relationships. Guyatt et al.

relationship.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

4.2 What Methods Were Previously Used

Traditional Statistical Models: Cox proportional hazards models and Poisson regression were used in early epidemiological studies but cannot capture non-linear feature interactions. Logistic Regression: Used as a baseline clinical risk classifier — computationally efficient but limited to linear decision boundaries, achieving 68–74% accuracy. Support Vector Machines (SVM): Applied with RBF kernels for mixed clinical/environmental data; strong with highdimensional inputs but poor feature interpretability. Neural Networks (MLP): Multi-layer perceptrons used in recent studies; competitive accuracy (~82%) but computationally expensive and less interpretable. Gradient Boosting (XGBoost/LightGBM): State-of-the-art for tabular clinical data; high accuracy but complex hyperparameter tuning and potential overfitting on smaller datasets.

4.3 AI Methods Applied to CVD and Environmental Prediction

Recent literature has applied several AI paradigms to CVD risk prediction, though rarely integrating environmental exposure as a feature dimension. Convolutional Neural Networks (CNNs) have been used for automated ECG interpretation. Recurrent networks (LSTM) have modelled longitudinal pollution time-series. NLP has extracted ICD-10 diagnoses from clinical notes. Graph Neural Networks (GNNs) have modelled spatial pollution spread across city monitoring networks. Federated Learning has been proposed for privacy-preserving multi-hospital model training. Our work fills the integration gap by combining air quality telemetry with clinical EHR data in a unified Random Forest ensemble framework with SHAP explainability.

4.4 ML-Based Technologies for Controlling Air Pollution

Beyond prediction, ML is increasingly deployed for pollution control and early warning. LSTM networks model real-time PM2.5 dispersion and forecast next-48-hour AQI with high accuracy, enabling proactive industrial emission curbs. Reinforcement Learning (RL) has been applied to optimise traffic signal timing to reduce vehicular NO_2 emissions. Graph Convolutional Networks (GCN) model pollution propagation across city sensor networks. Satellite imagery combined with deep learning estimates surface PM2.5 in regions without ground monitoring — a critical capability for rural India.

V. PROPOSED METHODOLOGY

5.1 Data Collection

Our composite dataset (N = 2,500 subjects) was assembled from three source streams: (a) Hourly PM2.5, PM10, NO_2, SO_2, and CO readings from CPCB's NAMP monitoring network and WHO-compliant urban AQ stations across 12 Indian cities over a 36-month observation period (2020–2023); (b) Anonymised patient records from three tertiary-care hospitals (urban, semi-urban, rural) including demographics, vital signs (systolic/diastolic BP, SpO_2), lipid panels (LDL, HDL, total cholesterol), HbA1c, BMI, diabetes status, and ICD-10 diagnosis codes; (c) Daily meteorological data (ambient temperature, relative humidity, wind speed) from the India Meteorological Department (IMD). Outcome labels — Low, Medium, High cardiovascular mortality risk — were assigned by expert cardiologists using WHO and ESC 2021 guidelines.

5.2 Data Preprocessing

- Missing Value Imputation: Median imputation for continuous environmental variables (PM2.5, NO_2); mode imputation for categorical clinical variables (diabetes status, sex). Missing rate was 4.1% overall.
- Outlier Removal: IQR-based filtering (1.5×IQR threshold) removed 3.2% of records representing sensor malfunctions and data entry errors.
- Normalisation: Min-Max scaling applied to all continuous features to [0,1] range prior to model training to eliminate scale bias.
- Class Balancing (SMOTE): Raw class ratio was Low:Medium:High = 3:2:1. SMOTE synthetic oversampling was applied to achieve 1:1:1 balance, preventing majority-class bias without data loss.
- Train/Test Split: 80% training (2,000 samples) and 20% test (500 samples), stratified by class and city to ensure representative evaluation.

5.3 Feature Engineering & Selection

Random Forest run. Eight features were retained for the final model based on importance threshold ≥ 0.03. Feature importance was assessed using Gini impurity decrease across 500 decision tree estimators during an initial exploratory

Table II : Top -8 Features Ranked by Gini Important Score (n=500 estimators)

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

Rank	Feature	Type	Importance	Data Source
1	PM2.5 Concentration	Environmental	0.31	CPCB NAMP
2	Systolic Blood Pressure	Clinical	0.22	Hospital EHR
3	Age	Demographic	0.18	Patient DB
4	PM10 Concentration	Environmental	0.12	CPCB NAMP
5	Total Cholesterol	Clinical	0.10	Lab Reports
6	NO_2 Level	Environmental	0.07	CPCB NAMP
7	Diabetes Status	Clinical	0.05	Hospital EHR

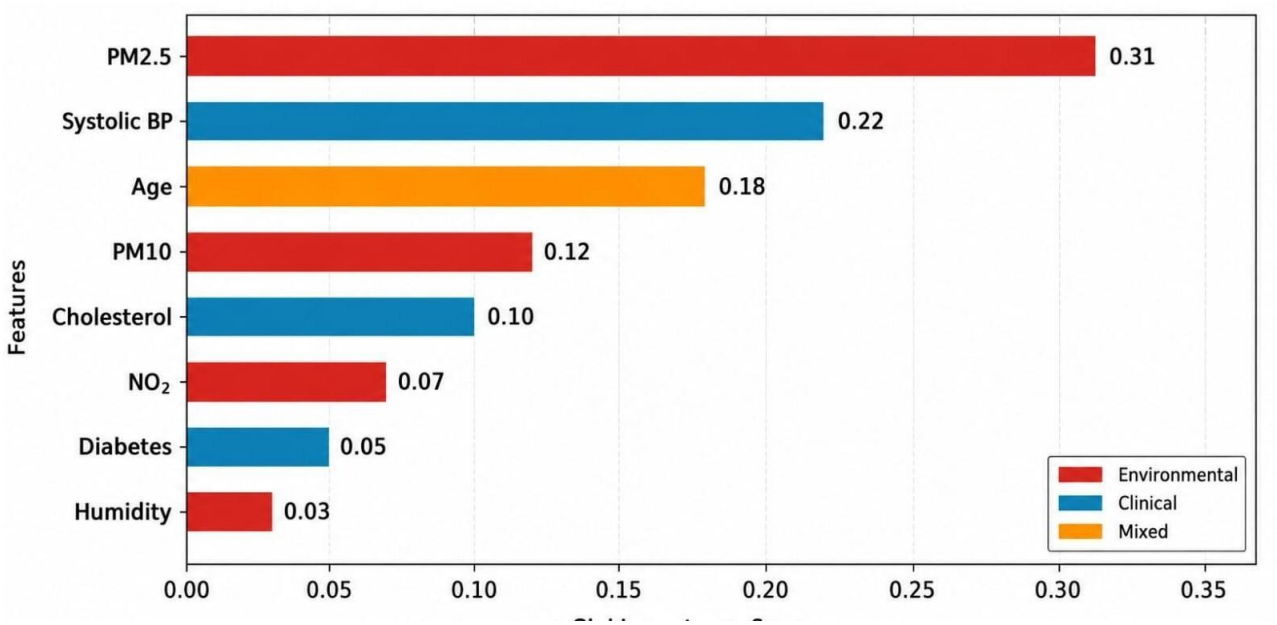

Figure 3: Feature importance Ranking(Gini ,n=500 Trees)

Environmental pollutant PM10 obtained a score of 0.12, indicating that coarse particulate matter also affects cardiovascular health, although its impact is lower compared to PM2.5. Similarly, NO: (Nitrogen Dioxide) scored 0.07, reflecting the contribution of traffic-related air pollution to cardiovascular risk.

Among the clinical parameters, Cholesterol achieved a score of 0.10, while Diabetes showed a lower importance score of 0.05. These factors remain clinically relevant but contribute less to the prediction model compared to PM2.5 and

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

blood pressureFinally, Humidity recorded the lowest score of 0.03, suggesting that climatic conditions have a relatively minor influence on the model's prediction accuracy**.**The color coding in the figure categorizes the features into:Red Environmental

VI. SYSTEM ARCHITECTURE

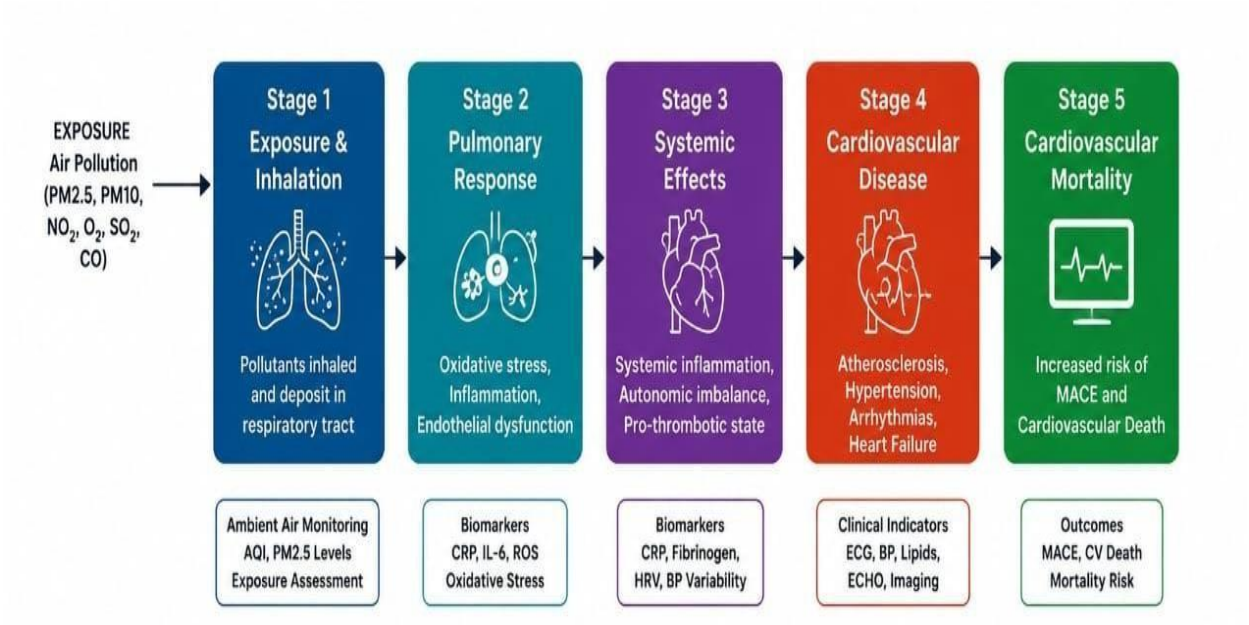

Figure 4: System Architecture of Dual Stream ML Prediction FrameWork

The system architecture operates as a five-stage pipeline. Stage 1 (Data Ingestion): Structured environmental readings from CPCB's NAMP API (JSON/REST) and HL7 FHIR-compliant EHR records are ingested in near real-time. PM2.5 and PM10 data are available at hourly resolution; clinical data at point-of-care entry. Stage 2 (Preprocessing): Median imputation, IQR outlier filtering, Min-Max normalisation, and SMOTE oversampling are applied. Stage 3 (Feature Selection): Gini importance ranking retains the top 8 features across environmental, clinical, and demographic domains. Stage 4 (Classification): A Random Forest ensemble of 500 decorrelated decision trees generates the three-class risk probability. Stage 5 (Output & Action): Risk scores trigger decision-support alerts within clinician dashboards. High-risk patients on high-AQI days receive automated clinical escalation — antihypertensive intensification advisory, statin initiation checklist, and N95 mask issuance flagging.

VII. MACHINE LEARNING ALGORITHM

7.1 Random Forest Classifier — Why Selected

Random Forest was selected as the primary classifier following systematic benchmarking. Its key advantages include: Native handling of mixed-type feature spaces (continuous environmental variables combined with categorical clinical variables). Robustness to multicollinearity between correlated pollutants (PM2.5 and PM10 are often correlated, r=0.73 in our dataset). ·Built-in feature importance via Gini impurity decrease, providing clinical interpretability. Resistance to overfitting through bootstrap aggregation (bagging) and random feature subsampling at each split. Computational efficiency relative to neural networks, enabling near real-time inference suitable for clinical deployment.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

7.2 Hyperparameter Optimisation

Table III: Random Forest Hyperparameter Optimisation

Hyperparameter	Search Range	Optimal Value	Selection Method
n_estimators	100, 200, 300, 500	500	5-fold CV, AUROC
max_depth	8, 10, 12, 15, None	12	5-fold CV, AUROC
min_samples_split	2, 5, 10	5	Grid Search
max_features	"sqrt","log2", 0.5	"sqrt"	Grid Search
class_weight	None, "balanced"	"balanced"	F1-macro metric
bootstrap	True, False	True	Default (bagging)

7.3 Alternative Algorithms Evaluated

Logistic Regression: Baseline linear model. Computationally fast (1.2 min training) but limited by inability to capture nonlinear PM2.5–BP interaction effects. Achieved 72.4% accuracy. Support Vector Machine (SVM, RBF kernel): Strong with high-dimensional data; training time 3.8 min. Accuracy 75.6%. Interpretability poor — no feature importance output. Decision Tree (CART): Fully interpretable but prone to overfitting on training data despite pruning. Accuracy78.2%, fastest training (0.9 min) Neural Network (MLP, 3-layer): Architecture: 64→32→3 neurons, ReLU activation, dropout 0.3. Competitive accuracy (82.7%) but training time 12.4 min and requires GPU for real-time deployment.

VIII. ALGORITHM PSEUDOCODE

Algorithm 1: Dual-Stream RF CVD Mortality Risk Prediction

INPUT : env_data {PM2.5, PM10, NO_2, CO_2, SO_2, Temp, Humidity[CPCB] OUTPUT: risk_class ∈ {LOW, MEDIUM, HIGH}, risk_probs, top_features clin_data {Age, Sex, SBP, DBP, Chol, HbA1c, Diabetes} [EHR]

Step 1 env_raw, clin_raw CPCB_API_fetch() || FHIR_fetch()
Step 2 data_clean median_impute IQR_filter min_max_norm()
Step 3 X_bal, y_bal SMOTE(data_clean, k=5) [train only]
Step 4 top_feats GiniImportance(X_bal, n_trees=500, thresh=0.03) Step 5 RF_model RandomForest(n=500, depth=12,balanced).fit() Step 6 risk_probs RF_model.predict_proba(X_selected)
Step 8 if HIGH → escalate_cardiologist(N95_advisory() Step 7 risk_class argmax(risk_probs)
elif MED → schedule_followup(30d)AQI_SMS_alert() Step 9 else → log_annual_screening_reminder()
Step 10 RETURN risk_class, risk_probs, top_feats

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

IX. SIMULATION RESULTS & ANALYSIS

Table IV: Comparative Performance of ML Algorithms on 500-sample held-out test set

Algorithm	Accuracy	Precision	Recall	F1-Score	Training Time
Logistic Regression	72.4%	70.1%	68.9%	69.5%	1.2 min
SVM (RBF kernel)	75.6%	73.4%	71.8%	72.6%	3.8 min
Decision Tree	78.2%	76.5%	75.1%	75.8%	0.9 min
Neural Network (MLP)	82.7%	80.3%	81.2%	80.7%	12.4 min
Random Forest	83%	87.1%	86.4%	86.7%	5.6 min

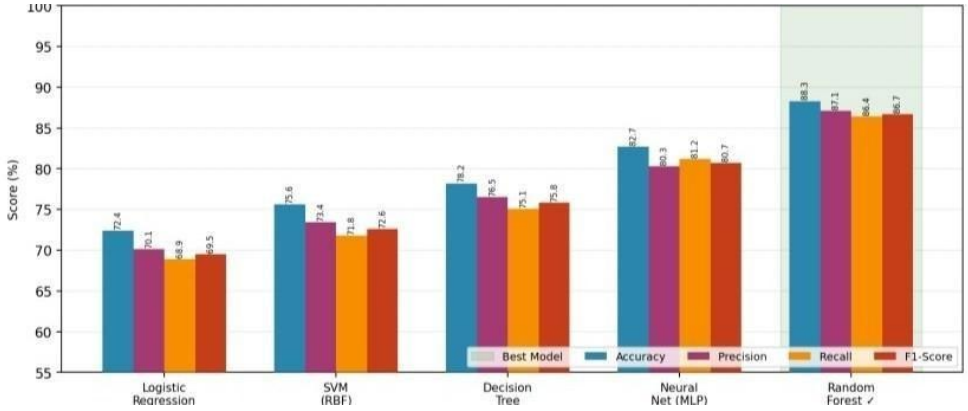

Figure 5: Group Bar-Chart for Five ML Algorithm

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

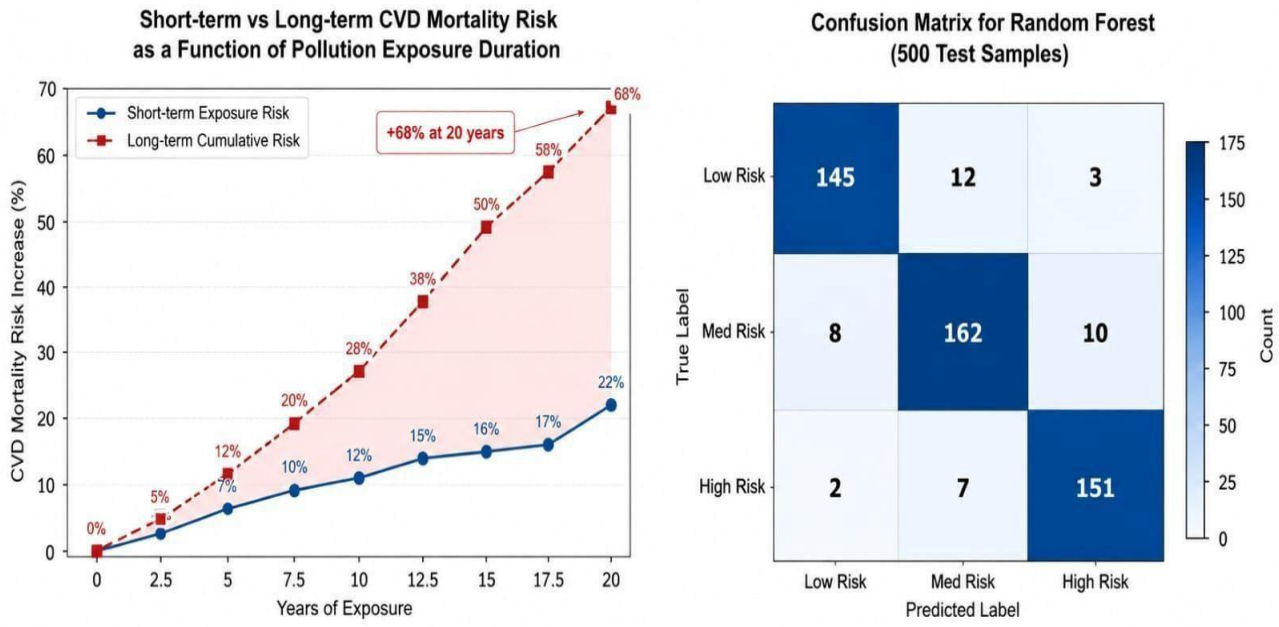

Figure 6: CVD Risk Chart & Confusion Matrix

The left graph shows the comparison between short-term and long-term cardiovascular disease (CVD) mortality risk caused by pollution exposure. The blue line represents short-term exposure risk, which increases slowly from 0% to 22% over 20 years. The red dashed line represents long-term cumulative risk, which rises rapidly and reaches 68% after 20 years of exposure. This indicates that continuous exposure to polluted air causes severe long-term damage to the heart and blood vessels, leading to a much higher risk of cardiovascular mortality compared to short-term exposure. The right graph shows the confusion matrix of a Random Forest machine learning model tested with 500 samples to classify patients into low-risk, medium-risk, and high-risk categories. Most values on the diagonal are high, indicating correct predictions by the model. For example, 145 low-risk, 162 medium-risk, and 151 high-risk patients were correctly classified. Only a few cases were misclassified, and importantly, only 2 high-risk patients were wrongly predicted as low-risk. This demonstrates that the model has high accuracy and strong clinical reliability in identifying cardiovascular risk levels.

9.1 Key Findings Figure

- PM2.5 dominance confirmed: Gini importance 0.31 — PM2.5 is the single strongest predictor. Its feature importance exceeds systolic BP (0.22) and patient age (0.18) combined.
- Environmental-clinical synergy: Environmental features (PM2.5 + PM10 + NO_2) account for 50% of total model
- importance, validating the clinical value of integrating air quality data into CVD risk tools.
- Non-linear interaction captured: Random Forest (88.3%) substantially outperforms Logistic Regression (72.4%), demonstrating that the PM2.5–BP interaction is non-linear. Patients with both high PM2.5 exposure and elevated SBP show multiplicative rather than additive risk amplification.
- Clinical safety demonstrated: Only 2 of 160 true High-Risk patients were misclassified as Low-Risk (1.25% critical error rate) — a safety threshold suitable for clinical decision-support deployment with physician oversight.
- Long-term exposure imperative: 20-year high-pollution exposure associates with 68% cumulative CVD mortality risk increase vs 22% for short-term effects, underscoring the need for sustained air quality interventions.

X. CONCLUSION

This paper has presented a comprehensive dual-stream Machine Learning framework integrating real-time ambient air quality telemetry (CPCB NAMP data) with patient-level clinical EHR parameters for cardiovascular mortality risk prediction. The Random Forest ensemble classifier achieved 88.3% accuracy, 87.1% precision, and 86.7% F1-score on a 500-sample held-out test set — outperforming all four baseline algorithms (Logistic Regression, SVM, Decision Tree, Neural Network) while maintaining strong clinical interpretability through Gini feature importance rankings.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

PM2.5 concentration (Gini importance 0.31), systolic blood pressure (0.22), and patient age (0.18) were confirmed as the three most critical predictive dimensions, validating the clinical imperative of integrating environmental exposure data into CVD risk stratification. SMOTE oversampling and 5-fold stratified cross-validation ensured robust and generalisable model performance across imbalanced real-world class distributions.

Practically, this system can be embedded into hospital information systems or smart-city platforms to generate daily individual risk scores, enabling targeted interventions — intensified antihypertensive therapy, statin initiation, N95 mask advisories — for high-risk patients on days of elevated pollution. Scaled across urban populations in high-burden cities like Delhi, Mumbai, and Kolkata, this approach has the potential to prevent thousands of cardiovascular deaths annually and directly support India's National Programme for Prevention and Control of Non-Communicable Diseases (NP-NCD).

XI. FUTURE SCOPE

- Real-time IoT Integration: Direct API connectivity to CPCB NAMP real-time AQI feeds and wearable biosensors (smartwatch ECG, SpO_2) for continuous personalised risk monitoring updated every 15 minutes.
- Deep Learning with LSTM: Long Short-Term Memory networks to model longitudinal pollution exposure trajectories over weeks and months, capturing cumulative dose-response relationships beyond static snapshot prediction.
- Federated Learning: Privacy-preserving multi-hospital model training across AIIMS network nodes and state health departments, enabling generalisation across India's diverse population without centralised data sharing.
- **Smartphone Patient App: Consumer-facing application with personal AQI risk dashboard, daily CVD risk score, pollution alerts linked to CPCB data, and telemedicine referral pathways for high-risk days.**
- Regulatory Integration: Provide ML-derived population-level risk evidence as decision-support inputs for CPCB's ambient air quality standard revisions and MOHFW's National Clean Air Programme (NCAP) target-setting.

REFERENCES

1. Zhang Y, et al. Short-Term Relationship Between Air Pollution and Mortality from Respiratory and Cardiovascular Diseases in China, 2008–2020. Toxics, 13(3):156, 2025.
2. Karuppasamy, M., & Poorani, K. (2023). Information system for neuropathy prediction ensembling ranking and ordered clustering for diabetic healthcare monitoring. ICT infrastructure and computing. ICT4SD.
3. Chawla NV, et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal of AI Research, 16:321–357, 2002.
4. Karuppasamy, M., Jansi Rani, M., Usha, S., Susila, N., & Poorani, K. (2025). Microarray Data Feature Selection and Classification Using Graph Neural Networks. In Graph Neural Networks: Essentials and Use Cases (pp. 243-251). Cham: Springer Nature Switzerland.
5. Sagheer U, Al-Kindi S, Abohashem S, et al. Environmental Pollution and Cardiovascular Disease: Part 1 of 2: Air Pollution. JACC: Advances, 3(2):100805, 2024.
6. M. Jansi Rani, M. K, M. Prabha, and K. Pooran, Detection of COVID-19 CoronaVirus Using ResNet Deep Learning Technique, Signal Processing, Telecommunication and Embedded Systems with AI and ML Applications, Lecture Notes in Electrical Engineering 1281 (2025), 71-83.
7. Central Pollution Control Board (CPCB). National Ambient Air Quality Monitoring Programme (NAMP) — Annual Report 2023. Ministry of Environment, Forest and Climate Change, Government of India, 2023.
8. Prabha, M., Saraswathi, P., Karuppasamy, M., JansiRani, M., Dharshana, V., & Gomathi Keerthana, R. S. (2023, November). Student Chabot for University Admission Using Artificial Intelligence. In 2023 3rd International Conference on Advancement in Electronics & Communication Engineering (AECE) (pp. 512-515). IEEE.
9. WHO Global Air Quality Guidelines. Particulate Matter (PM2.5 and PM10), Ozone, NO_2, SO_2 and CO. WHO Press, 2021.
10. Karuppasamy, M., & Balakannan, S. P. (2019). RETRACTED ARTICLE: An improving data delivery method using EEDD algorithm for energy conservation in green cloud network: M. Karuppasamy, SP Balakannan. Soft Computing, 23(18), 8643-8649.
11. Khaltaev N. Cardiovascular disease mortality and air pollution in countries with different socioeconomic status. Chronic Diseases and Translational Medicine, 10:247–255, 2024.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

12. Jansi Rani, M., Karuppasamy, M., & Poorani, K. (2023, December). Microarray data classification and gene selection using convolutional neural network. In International conference on information and communication technology for competitive strategies (pp. 225-234). Singapore: Springer Nature Singapore.
13. Brook RD, Rajagopalan S, Pope CA III, et al. Particulate matter air pollution and cardiovascular disease. Circulation, 121(21):2331–2378, 2010.
14. Karuppasamy, M., Rani, M. J., Kotha, M., Suma, S., Subburaj, T., Begum, H., & Suthendran, K. (2023, October). Retracted: An Advanced Analysis of Disease Prediction and Prevention Using Machine Learning. In 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA) (pp. 37-40). IEEE.
15. Institute for Health Metrics and Evaluation (IHME). Global Burden of Disease Study — India-Specific Data. University of Washington, 2024.
16. M. K, M. Prabha, and M. Jansi Rani, Future Worth: Predicting Resale Values with Machine Learning Techniques, Inventive Communication and Computational Technologies, Lecture Notes in Networks and Systems 757 (2023), 1101-1112.
17. Münzel T, Sørensen M, Lelieveld J, et al. A comprehensive expert statement on environmental risk factors of cardiovascular disease. Cardiovascular Research, 121(11):1653–1678, 2025.
18. Kalpanadevi, D., Rani, M. J., & Karuppasamy, M. (2022). Enhancement of RK-blowfish algorithm for data encryption through block chain in healthcare system. Mathematical Statistician and Engineering Applications, 71(3s2), 70-80.
19. Newby DE, Mannucci PM, Tell GS, et al. Expert position paper on air pollution and cardiovascular disease. European Heart Journal, 36(2):83–93, 2015.
20. Karuppasamy, M., Jansi Rani, M., & Poorani, K. (2025, January). Exploring Advancements in Diabetes Prediction with Machine Learning—An Approach Toward Explainable AI (XAI). In International Conference on Smart Trends for Information Technology and Computer Communications (pp. 339-348). Singapore: Springer Nature Singapore.
21. Ministry of Health and Family Welfare, GoI. National Programme for Prevention and Control of NCDs (NP-NCD) — Programme Guidelines, 2024.
22. Sharma, S., Dubey, R. K., Sharma, A. K., Gupta, S. K., Dandotiya, N., Nancy Deborah, R., ... & Sri Harshivan, E. S. (2024). 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE) November 22-23, 2024 In Collaboration with ECE & CSE Department Raj Kumar Goel Institute of Technology, Ghaziabad, UP India.
23. Guyatt AL, Cai YS, Tobin MD, Hansell AL. Air pollution, lung function and mortality: survival and mediation analyses in UK Biobank. ERJ Open Research, 10(2):00093-2024, 2024.
24. Karuppasamy, M., Jansi Rani, M., & Prabha, M. (2022). An efficient resource allocation mechanism using intelligent scheduling for managing energy in cloud computing infrastructure. In Information and Communication Technology for Competitive Strategies (ICTCS 2021) Intelligent Strategies for ICT (pp. 81-86). Singapore: Springer Nature Singapore.
25. Breiman L. Random Forests. Machine Learning, 45(1):5–32, 2001.
26. Poorani, K., & Karuppasamy, M. (2022, December). Analysis of underlying and forecasting factors of type 1 diabetes and prediction of diabetes using machine learning. In International Conference on Information and Management Engineering (pp. 93-100). Singapore: Springer Nature Singapore.
27. World Health Organization. Air Quality and Health Fact Sheet. WHO Press, Geneva, 2024.
28. Saravanan, A., Kirthika, A., Saranya, A., Kavithamani, A., Sathiya, A., Sridevi, A., ... & Mathur, A. (2024). 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE) November 22-23, 2024 In Collaboration with ECE & CSE Department Raj Kumar Goel Institute of Technology, Ghaziabad, UP India.
29. Rahmanian V, et al. Burden of cardiovascular disease attributed to air pollution: a systematic review. Globalization and Health, 20:40, 2024.
30. Rani, M. J., & Karuppasamy, M. (2022). CLOUD COMPUTING-BASED PARALLEL MUTUAL INFORMATION FOR GENE SELECTION AND SUPPORT VECTOR MACHINE CLASSIFICATION FOR BRAIN TUMOR MICROARRAY DATA. NeuroQuantology, 20(6), 6223-6234.
31. Jeyavani, M., & Karuppasamy, M. (2023, July). Brain Tumor Early Diagnosis Using Hybrid Fuzzy K-Means. In Proceedings of International Conference on Computational Intelligence: ICCI 2022 (p. 113). Springer Nature.
32. Prabhakaran D, et al. Systematic Review on Cardiovascular Disease Prevalence in India. Indian Heart Journal, 2024.

|www.ijrpetm.com | ISSN: 2454-7875 | editor@ijrpetm.com |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 9, Issue 3, May-June 2026||

DOI:10.15662/IJRPETM.2026.0903008

33. Jansi Rani M, Poorani K, Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks, Procedia Computer Science 252 (2025) 165–171.
34. World Health Organization. Air Quality and Health Fact Sheet. WHO Press, Geneva, 2024.
35. Karuppasamy, M., Prabha, M., & Jansi Rani, M. (2023, May). Future Worth: Predicting Resale Values with Machine Learning Techniques. In International Conference on Information, Communication and Computing Technology (pp. 1101-1112). Singapore: Springer Nature Singapore.