International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.157

Temporal Air Quality Forecasting using Hybrid RNN Model.

DharsanGM
DepartmentofDataScienceandBusinessSystems,
SchoolofComputing,
CollegeofEngineeringandTechnology,
SRMInstituteofScienceandTechnology,
Kattankulathur,Chennai,603203,India.

PriyadarsiniK
DepartmentofDataScienceandBusinessSystems,
SchoolofComputing,
CollegeofEngineeringandTechnology,
SRMInstituteofScienceandTechnology,
Kattankulathur,Chennai,603203,India.

***Abstract*—Inmanynations,poorairqualityhasturnedintoa significant environmental issue. The majority of air pollutantsseriouslyharmpeople'shealthandqualityoflife.Rapidurbanisation,industrialexpansion,andtrafficareallcontributingto a significant decline in air quality. In this work, we suggest ahybrid recurrent neural network (RNN) model that combines thebenefits of wavetransform decomposition and LSTM model fortemporalairqualityforecasting.Thetimeseriesisdividedintoitstrend, seasonal, and residual components using the decompositiontechnique, which is then fed into the LSTM model for forecasting.Weexaminedtheeffectivenessofthesuggestedmodelusingdailyairqualitydatacollectedfroma monitoring station in Delhi.OuroutcomeindicatethattheH-LSTMmodeloutperformstraditionaltime-series models and the LSTM model in termsof predicting accuracy, having reduced the mean absolute errorandmeansquarederrorof50%forPM2.5concentrations, performs better in terms of forecasting accuracy than traditionaltime-series models and the Long short - term memory model.Furthermore, the decomposition technique allows us to identifytheunderlyingtrendsandseasonalpatternsinthedata,whichcanprovideinsightsforairqualitymanagement.Ourresults highlight the potential for further advancements in thisfield andshow the value of using a decomposition technique and LSTMmodelforairqualityforecasting.**

***IndexTerms*—Airquality,wavedecomposition,PM2.5concentration,LSTM**

I. INTRODUCTION

On a global level, air pollution poses a serious hazard tohumanhealththatisgrowinginimportance.TheWorldHealth Organization (WHO) 2018 report states that 92% of peopleworldwidebreathetoxicair,whichisthoughttobethecauseof7 million annual fatalities. Two respiratory and cardiovasculardisorders,emphysemaandasthma,areamongthe harmfulimpacts of pollution in the air on human health. In addition,ambientairpollutionhas1.8yearsofftheaveragelifetime.

[9] The World Bank estimates that the economic costs of airpollution, including lost productivity, medical costs, and earlydeaths,total$5trillionannually.By2050,itisanticipatedthat the number of premature deaths brought on by particulatematter(PM)willhavemorethantripled.Thus,itisessentialto identify a precise and practical method to lower ambient airpollution if we are to guarantee the security and well-being oftheglobalpopulace.

Wecannotseepollutants,andwearenotawareofthereasons whypollutionlevelsarerising.Tocomprehendtheorigins,itisnecessaryto firstgooverthe principlesofairpollution[3].PM2.5endangerstheenvironmentglobally.Long-termexposuretotheseparticles,whichcanquicklyreachdeepwithintherespiratoryorgans,canresultinavarietyofdiseases.ThediameterofPM2.5,whichismuchsmallerthanthediameterofahumanhair,is2.5microns.Asaresult,thereis asignificant sourceof exposureand sicknessonthe humanbody.Theyendangerhumanhealth,raisingthelikelihood ofcancerandotherendemicandpandemicdiseases.Theultimate goalistoraisepublicawarenessofthevariablesthataffectenergyconsumptionandmethodsforcontrollingit.[8]

InthispaperwepresentahybridRNNmodelthatcombinesthedecompositiontechniquewithlongshort-termmemoryfor temporal air quality forecasting (LSTM). While the de-compositiontechniquesareusedtobreakdownthetimeseries into wave transforms, which aid in identifying trends,cyclic patterns, etc., the LSTM cells are renowned for theircapacity to capture long-term dependencies in the data. Theproposed model also incorporates input features, which havebeen demonstrated to be significant predictors of air quality,suchasmeteorologicalvariablesandtimeofday.

Usingactualairqualitydatagatheredfromamonitoringstationin Delhi,weassessedtheperformanceofthesuggestedmodel. Withameanabsoluteerrorofg/m3forPM2.5concentrations,our resultsshowthatthehybridRNNmodeloutperformsconventionaltime-seriesmodelsandotherRNNmodelsintermsofforecastingaccuracy.Inaddition,thehybridRNNmodeloffersusefulinsightsintothetemporaltrendsinairpollutionthatmaybeutilisedtoguidepolicyanddecision-makingforthemanagementofairquality.Therestofthispaperisstructuredasfollows.AsummaryofrelatedresearchemployingRNNmodelsforairqualityforecastingisincludedinSection2.ExperimentalareaofstudyexplainedinSection3.Theexperimentalsetupandproposedmethodologyisexplainedinsection4.The findingsofourmodelevaluationareshowninSection5and6.Section7concludesthepaperandconsiderspossiblefuturelinesofenquiryforthis fieldofstudy.

II. RELATEDWORKS

DOI: rp-9788770040723.157

In order to forecast particulate matter (fraction PM2.5) airpollution, a study comparing artificial neural networks (ANNs)andadaptiveneuro-fuzzyinferencesystems(ANFIS)wasconducted. The results are provided in this publication. PM2.5hourly measurement records from the Airbase databases wereused for the trials. The two key statistical measures calculatedwere mean absolute error and root mean square error (MAE).[6]

This study proposes improved models for estimating hourlyair pollution concentrations using meteorological data frompreviousdays.Amulti-tasklearning(MTL)problemwithvariousregularisationtechniquesisusedtoformulatethepredictionoverthenext24hours.Theresultsshowthatthe parameter-reducing formulations and regularisations as-sociatedwithsuccessivehoursoutperformcommonlyusedstandard regression models and regularisations for MTL whencomparedtotheproposedregularisation.[10]

Thisstudyusesalongshort-termmemory(LSTM)approachtoforeseeO3,PM2.5,NOx,and COconcentrationsinDelhi,theheavilypollutedIndianNational CapitalTerritory.Theresearchadoptsiveseparatesetsofparametersandcriteria,suchastrafficdata,levelsofpollution,climaticstatus,andvehicleemission.TheLSTMmodelsaccuratelyestimate hourlyconcentrationswhilehandlingthecomplexitiesoflong-terminteractionscomingfrombothhumanandnaturalsources.Theresultsofthisstudymaybeusedbythegovernmentandpolicymakerstoplanactionstoreducethenegativeeffectsofdecliningairquality.[4]

Becausepollutantscanbeharmfultohumanhealth,predictingairpollutionisahottopic.Traditionalmachinelearningmodelstypicallyfocusonincreasingpredictionaccuracyoverallwhileignoringaccuracyforpeakvalues.Furthermore,itisimpossibletocomprehendthesemodels.Theydonotadequatelyexplainhowvariousdecidingfactorsinteracttoaffectairpollution.Inthisstudy,anewcombinationofinterpretablepredictivemachinelearningmodelisenhancedwithtwonoveltiesforthepredictionofPM2.5.First,afusionmodelframeworkisdevelopedusingdeepneuralnetworksandanonlinearautoregressivemovingaveragewithexogenousinputmodel.Second,thishybridmodelincludesmethodsfor automaticallycreatingandchoosingfeatures.Theexperimental findingsshowthat ourmodeloutperformsothermodelsintermsofhighestvaluepredictionprecisionandmodelinterpretability.TheproposedmodelillustrateshowtocalculatePM2.5estimatesusingexistingPM2.5,climate,andseasondata.Therecommendedapproachalsoprovidesanoveleasilyunderstoodmachinelearningframeworkfortemporalseriesdata,enablingthecreationofprecisepredictivemodelsandthedefenceofintricaterelationshipsbetweenmultimode inputs.[2]

Globally,airpollutionisakeyfactorinprematuremortality.It’s crucial to understand and anticipate air pollution patternsto lessen its caused damage. Complex prediction algorithmsareneededforthis.InordertounderstandPM2.5patternsthroughout space and time, deep learning models like GCNNand ConvLSTM are used. Remote sensing satellite photos areutilised to monitor atmospheric pollutant issues while time-series multidimensional directed graphs are employed to char-acterise climate aspects. ConvLSTM is used to foresee PM2.5in Los Angeles region using data from the preceding 10 daysandground-basedPM2.5sensordata.Thespatiotemporaldeeppredictive algorithms outperform earlier studies significantly.[5]

Intheproposedstudy,ahybridapproachisintroducedfor effectively breaking down single, one-dimensional PM2.5time data into various dimensional time data and extractinghiddeninformationfromit.Eachsequenceispredicted bythealgorithmusingtraditionalpredictiontechniques,andthe finalpredictionsareobtainedbyreconstructingtheresults.Three hybridmodelsW-ANN,W-ARIMA,andW-SVMweredevelopedinthestudytoforesee2016PM2.5trendsinfiveChinesecities.Accordingtotheoutputs,combinedmodelsperformbetterthantraditionalARIMA,ANN,andSVMmodelsatforecastingshort-termPM2.5concentrations.WhenitcomestoforecastingPM2.5 concentrations,theW-ARIMAmodelperformsbetter,especiallywhenitcomestocapturingthemutationalpointsthatcouldhelpcreatepollutionwarnings.[1]

A new model for predicting PM2.5 concentrations for eachhour during heavy haze episodes is presented in the proposedwork. For increased forecasting accuracy, this model com-bines the mode decomposition-recombination method with anensemble learning strategy. The data is divided into severalfrequency modes using the FEEMD to lessen the effects ofnoise in the data. Then, to ensure precise extraction of information and computational efficiency, related modes are combined and excessive breakdown is avoided using the sampleentropy (SE). As a forecasting model,SDEM is built which isstack driven to improve feature representation and informationconsumptioncapabilities.Eachbasemodel’sgeneralisationperformance is enhanced via K-fold cross-validation, and themeta-model generates higher prediction results by using eachbasemodel’soutputsasnewinputs.Regardingpredictingprecision, stability, and class prediction rate , the FEEMD-SE-SDEM model performs better than earlier contrast models,making it a suitable choice for an early air quality warningsystem.[7]

III. StudyArea

A. DataCollection

Delhi,thecapitalofIndia,islocatedinthenorthernpartof the nation.It is a big, multiethnic city that has a thrivingmodern culture and history. The city, which has a populationofaround20million,isahubforbusiness,politics,and education. Due to the large population, transportation, urbanisation,industrialization, etc. are all growing quickly. These are someof the factors that contribute to Delhi’s high pollution levels,whichcanimpairpeople’shealth.Thereare46livemonitoringstationsforairqualityinDelhi,oneofwhichIhavechosentobeinAshokVihar.ItisaresidentialneighbourhoodinDelhi’s northwest. It is roughly at latitude 28.6958°. Delhi hasa hot semi-arid climate, with extremely hot summers and coolwinters.Thetemperaturecanreachashighas45°C(113°F)in the summer months of May and June, while the wintertemperaturescanfallaslowas2°C(35.6°F)inDecembera

International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.157

nd January. The city experiences monsoon rains from July toSeptember, which provide relief from the hot summer months.Overall, the climate of Delhi is characterized by its extremetemperaturevariationsandseasonalwinds.

Fig.1.GeographicallocationofAshokVihar

B.DatasetDescription

The dataset collected from Ashok Vihar contain 1450 datawith11featureswhichincludeconcentrationslikePM2.5,PM10, NO2, NOxetc and meteorological factors like wind-speed, humidity, temperature, wind direction etc. The data iscollectedfromyear2019January-2022December.

IV. Proposed Methodology

The proposed methodoly is a technique which illustrates themodel used in this work step by step. The dataset is collectedfromhttps://cpcb.nic.in/.InthatwebsitefromAshokViharliveairqualitymonitoringstationthedataiscollectedforpast 4years.TheworkFowisprovidedbelowasshownin[?].The data is extracted in CSV format. Data preprocessing is atechniquewhichisusedtoprepareabetterdatafortrainingthemodel.Thenullvaluesinthedatasetareidentifedandreplacedwith distribution of each feature. For outlier detection a robustoutlierdetectionalgorithmHampelIdentifierisused.

A.HampelIdentfier

Toexcludeoutliersfromadataset,onetechniqueistoimplementtheHampelidentifier.Itstartsbyfiguringoutthedata'smedianandmedianabsolutedeviation(MAD).Bycomputingthemedianoftheabsolutedeviationfromthemedian,theMADameasureofthedata'svariabilityisestablished.Eachdatapointisthencomparedtothemedian;iftheabsolutedifferenceexceedsacertainnumber(oftenamultipleoftheMAD),itisdeemedanoutlierandeliminatedfromthedataset.Instatisticalanalysis,datamining,andmachinelearningapplications,theHampelIdentifierisuseful,especiallywhenworkingwithdatasetsthathaveerrorsorextremevalues.

B. Decomposition

Decompositionistheprocessofbreakingdowntimeseriesintowaveformatusingmathematicaltechniques.Itisusedtoidentifythehiddentrends,patternandcyclesintimeseriesdata.Therearedifferenttypesofdecompositiontechnique,EnsembleEmpiricalModeDecomposition(EEMD)isusedinthismodelfordecompsingthePM2.5feature.ThedecomposedPM2.5areshownin figure2.Sincemodemixingmakesitchallengingtodistinguishbetweenseveraloscillatorymodesinthedata,theEEMDapproachisanimprovementovertheEMDmethod.ThebreakdownbecomesmorediversewhennoiseisaddedtothedatA,makingiteasiertoidentifythemanyoscillatorymodes.ThismethodoffersastrongtoolforunderstandingtheunderlyingprocessesthatgeneratethedatA,makingitvaluable.ThePM2.5dataisbrokedownintoIMFsandresidual.TheIMFsarehelpfulincapturingthehighertrendsorhigherfrequencyandresidualusedtocapturethelowertrendsorlowerfrequencyinthemodel.

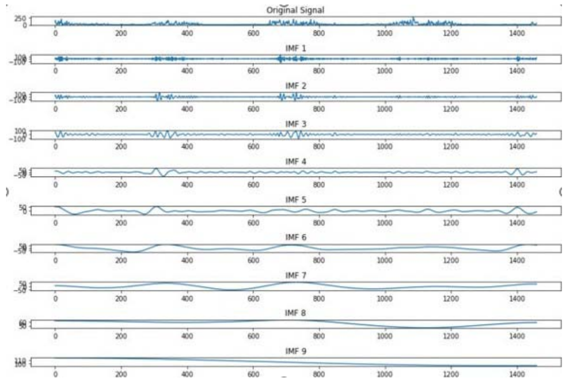

Fig.2.DecompositionofPM2.5

C.LSTM

TheissueoftypicalRNNs'vanishingorexpandinggradientsisresolvedbytheRNNarchitectureapproachknownasLSTM.Becauseitcanidentifylong-termdependenciesinthedata. LSTM is useful for time series data. LSTMs use memorycells to store information over time and gates made of sigmoidandtanhfunctionstoregulateinformationfow.Thesegatesincreasethemodel'sabilitytorepresentcomplexsequencesbyallowingittochooseretainorrejectinformationbasedontheinput data.Overall,LSTMhasbeenextensivelyappliedinmanydifferentapplicationsandisastrongtoolformodellingsequentialdata.Beforepassingthedatatothelstmodel,thedataisscaledwithfeature rangeof(o,1).LongShort-TermMemory(LSTM)networksusethreedifferenttypesofgates:input,output,andforgetgatesasshowninfgure3.InorderforLSTMstomanagelong-termdependenciesinsequentialdata,thesegatesarecrucialbecausetheyregulatethevowofinformationintoandoutofthememorycell.

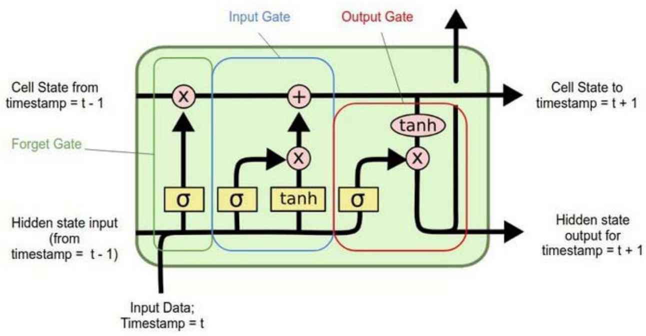

Fig.3.ArchitectureofLSTM

D.ModellingProcessofHybridLstmApproach

Overall,thisworkoffersahybridLSTM(HI-EEMD-LSTM)modelforthedailyPM2.5forecasting,followingthetech

DOI: rp-9788770040723.157

nicalpathof"decompositionandLSTM."Themodellingprocess isdepictedinfigure4,anditmaybesummedupasfollows:

Step-1:Gethistoricalinformation,suchasPM2.5concentrationsandweatherinformation.

Step-2:ToobtainnnumberofIMFsandoneresiduemode,performtheEEMD

Step-3:Standardizetheresultingdecomposedwavesinthefeaturerangefromthemeteorologicaldatafrom0to1

Step-4:Createa(n+1)LSTMmodelwitheachIMFacquiredfromdecompositionasthetrainingdatafortheweather.

Step-5:ToacquiretheoriginalPM2.5concentrations,dothe inverse EEMD to produce the anticipated PM2.5 concentrations

Step-6:Producetheforecastingexercise'sfinaloutput.

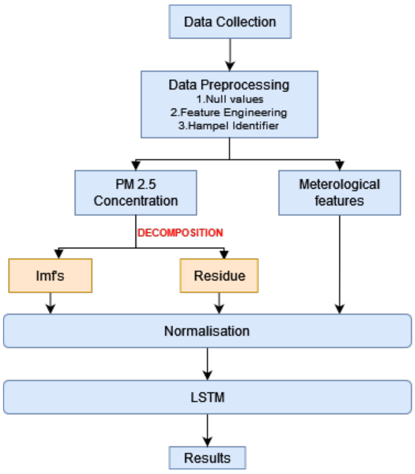

Fig.4.ArchitectureofHybridLSTMmodel

V. EvaluationMetrics

Many evaluation metrics root mean squared error (RMSE),mean absolute error (MAE), and mean squared error (MSE)areusedtoidentifythecapabilityofthemodel.Theevaluationmetrics of the model is compared with basic LSTM model toidentify how much better the proposed model is executingbetterthantheLSTMmodel

VI. Results

Thestudy's findingsdemonstratedthatthehybridLSTMmodelgreatlyperformedbetterthantheLSTMmodelsinpredictingPM2.5valuesbasedonmeteorologicaldata.TheMAEofthebestLSTMmodelwas 42.20,whichwasasignificantimprovementoverthebaselineMAEof91.83.Similarly,therootmeansquarederror(RMSE)wasalsosignificantlylowerfortheLSTMmodelscomparedtothebaselinemodels.TheLSTMmodelswereparticularlyeffectiveincapturingthelong-termdependenciesinthedata,whichhelpedtoimprovepredictions.Thefigure5andfigure6givesagraphicalrepresentationofboth";themodelswithactualandpredictedvaluesThese findingssuggestthatLSTMmodelsmaybeausefultoolforpredictingstockprices,andcouldhaveimportant applications in finance and investment. However, itisimportanttonotethattherewerebsomehindrancetothestudy,includingarelativelylittlesamplesizeandalimitedtimeperiodfortheanalysis.Furtherresearchisneededtovalidatethese findingsandexplorethepotentialofLSTMmodelsforotherapplicationsin financeandbeyond.

TABLEI: Model Performance

	EvaluationMetrics		
Model	MAE	MSE	RMSE
LSTM	91.83	12161.98	110
HI-EEMD-LSTM	42.2	3997.97	63.22

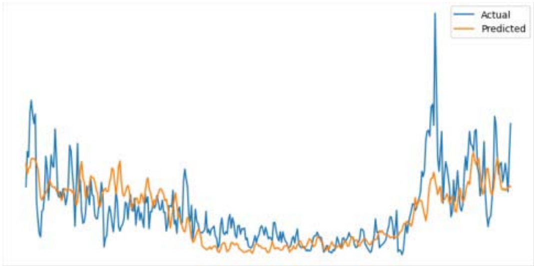

Fig.5.LSTMmodel

VII. Conclusion

Inthisstudy,aninnovativeHybridLSTMmodelisintroducedforpredictingdailyPM2.5concentrationsbyutilisingmultiple LSTMmodelsindifferentmodes.ThemodelfirstappliesEEMDtotransformthetimeseriesfusedfeaturesintosimplerfeaturesinmono-mode.Thematchingbetween

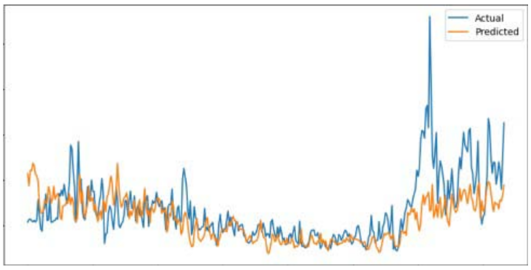

Fig.6.HybridLSTMmodel

meteorologicalconditionsandeachmode'scoefficientsisthenconstructed using LSTM, resulting in the development of agroup of LSTM models for ensemble learning. The predictedmode coefficients are then incorporated into the results usingthe reverse EEMD. Using wave decomposition, the H-LSTMmodelhasincreasedtheaccuracyofPM2.5observation. WhenthesuggestedH-LSTMmodeliscomparedtotheLSTMmodel on the Ashok Vihar dataset, it is evident that the H-LSTMmodeloutperformsbothofthosemethods.

References

[1] YongCheng,HongZhang,ZhenhaiLiu,LongfeiChen,andPingWang, "Hybrid algorithm for short-term forecasting of pm2. 5 in china,"Atmosphericenvironment,vol. 200, pp. 264–279,2019.

DOI: rp-9788770040723.157
[2] YuanlinGu,BaihuaLi,andQinggangMeng, “Hybridinterpretablepredictivemachinelearningmodelforairpollutionprediction,” Neurocomputing,vol. 468, pp. 123–136,2022.

[3] ChiouJyeHuang,andPing-HuanKuo,“ Adeepcnn-lstmmodelforpar-ticulatematter(pm2.5)forecastinginsmartcities,” Sensors,vol. 18, no. 7, p. 2220,2018.

[4] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ...&Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis.Computers and Electrical Engineering, 106, 108556..

[5] PratyushMuthukumar,EmmanuelCocom,KabirNagrecha,DawnComer, IreneBurga,JeremyTaub,ChisatoFukudaCalvert,JeanneHolm,andMohammadPourhomayoun,“ Predictingpm2.5atmosphericairpollutionusingdeeplearningwithmeteorologicaldataandground-basedobservationsandremote-sensingsatellitebigdata,” AirQuality,Atmosphere&Health,vol. 15, no. 7, pp. 1221–1234,2022.

[6] MihaelaOprea,SandaFlorentinaMihalache,andMarianPopescu, “Acomparativestudyofcomputationalintelligencetechniquesappliedtopm2.5airpollutionforecasting,”In20166thInternationalConferenceonComputersCommunicationsandControl(ICCCC),IEEE, pp.103–108,2016.

[7] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[8] Nur’atiahZaini,LeeWoenEan,AliNajahAhmed,andMarlindaAbdulMalek,“ Asystematicliteraturereviewofdeeplearningneuralnetworkfortimeseriesairqualityforecasting,” EnvironmentalScienceandPollutionResearch,pp.1–33,2021.

[9] GuangjiZheng,HuiLiu,ChengqingYu,YeLi,andZijieCao, “Anewpm2.5forecastingmodelbasedondatapreprocessing,reinforcementlearningandgatedrecurrentunitnetwork,” AtmosphericPollutionResearch,vol. 13, no. 7, p. 101475,2022.

[10] Dixian Zhu, ChangjieCai, Tianbao Yang, and Xun Zhou, “A machinelearning approach for air quality prediction: Model regularization andoptimization,” Bigdataandcognitivecomputing,vol. 2, no. 1, p. 5,2018.