

П.А.Кожобекова*, М.А. Сыздыкова, О.М. Сүлеймен, Ж.Д. Изтаев

т.ғ.к., доцент, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан

магистрант, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан

магистр, оқытушы, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан

п.ғ.к., қауымдастырылған профессор, М.Әуезов атындағы ОҚУ, Шымкент, Қазақстан

*Корреспондент авторы: pernesh-63@mail.ru

ҮЛКЕН ДЕРЕКТЕРДІ ӨНДЕУ АРҚЫЛЫ АТМОСФЕРАЛЫҚ ЛАСТАНУДЫ НАҚТЫ УАҚЫТ РЕЖИМІНДЕ МОНИТОРИНГТЕУ ЖҮЙЕСІН ҚҰРУ

Түйін

Мақалада үлкен деректерді (Big Data) өңдеу технологияларын пайдалану арқылы атмосфералық ауаның ластануын нақты уақыт режимінде мониторингтеу және болжау жүйесін құру мәселелері қарастырылады. Жүйенің деректерді жинау, тасымалдау, сақтау, өңдеу және визуализациялау кезеңдерінен тұратын архитектурасы сипатталады. Сенсорлар арқылы алынған экологиялық деректерді өңдеуде Apache Kafka және Apache Spark секілді бөлінген есептеу платформаларының тиімділігі талданады. Сонымен қатар машиналық оқыту алгоритмдерін қолдану арқылы ауа сапасының өзгерісін алдын ала болжау мүмкіндіктері LSTM, Random Forest, XGBoost, LightGBM және гибриді CNN-RNN модельдерінің нәтижелері негізінде көрсетіледі. Жұмыста экологиялық мониторинг жүйелерінің қоршаған ортаның жай-күйін бақылаудағы маңызы айқындалып, атмосфералық ластанудың адам денсаулығына әсерін азайтудағы рөлі қарастырылады. Қазақстан жағдайында мұндай интеллектуалды жүйелерді енгізудің экологиялық қауіпсіздікті арттырудағы, деректерді жедел талдаудағы және басқарушылық шешімдер қабылдаудағы практикалық маңыздылығы негізделеді.

Кілттік сөздер: үлкен деректер, атмосфералық ластану, нақты уақыттағы мониторинг, машиналық оқыту, PM2.5, Apache Spark, Apache Kafka.

Кіріспе. Атмосфералық ауаның ластануы адам денсаулығына, экожүйеге және климатқа тікелей әсер ететін қазіргі заманғы экологиялық қауіпсіздіктің өзекті мәселелерінің біріне айналып отыр. Әсіресе өнеркәсіптік аймақтар мен мегаполистерде ауа сапасының төмендеуі экологиялық қауіптердің артуына себеп болады. Ауадағы PM2.5, PM10, CO, SO₂ және NO₂ секілді зиянды заттардың концентрациясы тыныс алу жүйесі ауруларының көбеюіне ықпал етеді [1]. Қалалық ортадағы атмосфералық ластанудың негізгі көздеріне жылу-электр станциялары, өнеркәсіптік кәсіпорындар, автокөлік ағыны және тұрмыстық қыздыру жүйелері жатады.

Қазақстан Республикасының ірі қалаларында географиялық орналасуы, температуралық инверсия және өндірістік қызметтің әсерінен қыс мезгілінде PM2.5 концентрациясы Дүниежүзілік денсаулық сақтау ұйымы (ДДҰ) ұсынған шектен бірнеше есе асып түседі. Мысалы, Астана қаласында бұл көрсеткіш ДДҰ нормаларынан 19,8 есеге, ал Қарағанды қаласында 20,5 есеге дейін асқан жағдайлар тіркелген [2]. Мұндай жоғары ластану халықтың денсаулығына ауыр салдарлар әкеліп, тыныс алу жүйесі мен жүрек-қан тамыры ауруларының өсуіне себеп болады.

Дәстүрлі мониторинг жүйелері негізінен мемлекеттік стационарлық станцияларға негізделген, олардың саны шектеулі және кеңістіктік қамтуы төмен. Бұл олқылықтарды жою үшін соңғы жылдары бұлттық есептеулер мен Big Data технологиялары қолданысқа енгізілуде. Бұл шешімдер ауа сапасы туралы деректерді жоғары ажыратымдылықпен, нақты уақыт режимінде жинауға және өңдеуге мүмкіндік береді [3]. Қазіргі заманғы зерттеулерде атмосфералық ластануды бағалаудың бірнеше тәсілі ұсынылған: спутниктік деректерді қолдану, физикалық-химиялық модельдеу және сенсорлық желілерге негізделген бақылау.

Дегенмен осы тәсілдердің ішінде нақты уақыт режиміндегі мониторингте Big Data технологияларын біріктіру ең перспективалы бағыт ретінде көрсетіліп отыр, өйткені Big Data платформалары үлкен көлемдегі деректерді кідіріссіз өңдеуге мүмкіндік береді [4].

Осы мақаланың мақсаты - атмосфералық ластануды нақты уақыт режимінде мониторингтеу үшін Big Data технологиялары мен машиналық оқыту әдістерін біріктіретін жүйенің құрылымдық моделі мен компоненттерін талдау, оның негізгі қабаттарын анықтау және болжау дәлдігін бағалау бойынша ұсынымдарды қалыптастыру болып табылады.

Материалдар мен әдістер.

Үлкен деректерге негізделген нақты уақыттағы мониторинг жүйесінің архитектурасы өзара байланысты төрт негізгі қабаттан тұрады: деректерді жинау, тасымалдау, өңдеу және визуализациялау. Әрбір қабат белгілі бір функционалдық міндетті орындайды және жалпы жүйенің кідіріссіз әрі сенімді жұмыс істеуін қамтамасыз етеді. Зерттеуде осы архитектураны құрайтын компоненттер мен технологиялар салыстырмалы түрде талданады.

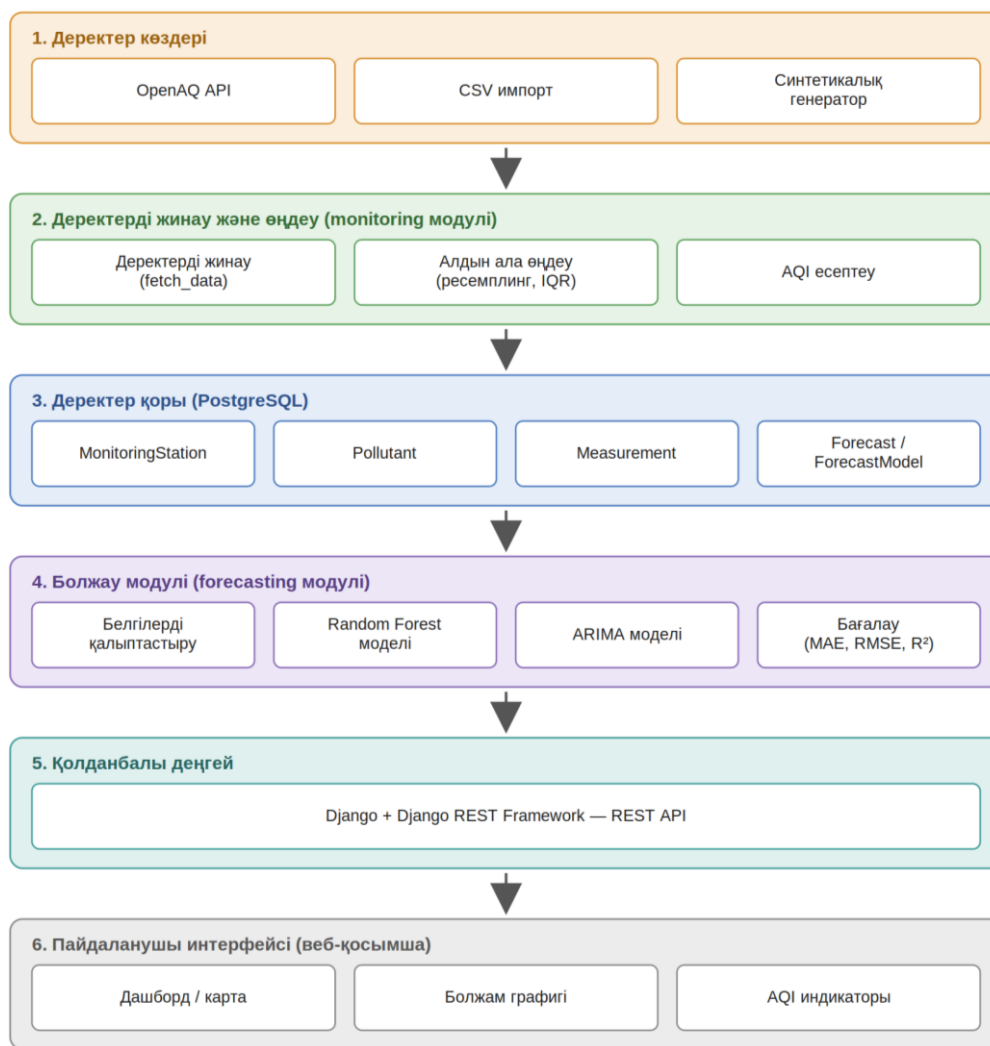
Деректерді жинау қабаты. Бұл деңгейде қала аумағында орналастырылған сенсорлар PM2.5, PM10, CO, NO₂, SO₂, O₃ сияқты ластаушы заттардың концентрациясын, сондай-ақ температура, ылғалдылық және атмосфералық қысым секілді метеорологиялық параметрлерді өлшейді. Соңғы зерттеулерде ESP8266 NodeMCU, Raspberry Pi және Arduino секілді микроконтроллерлерге негізделген арзан, бірақ тиімді сенсорлық құрылғылар жасалған. Олар Wi-Fi немесе GSM желілері арқылы өлшеу деректерін орталық серверге жібереді.

Деректерді тасымалдау қабаты. Сенсорлардан түскен ағынды деректерді өңдеу үшін Apache Kafka секілді брокерлік платформа қолданылады. Kafka жоғары өткізу қабілеті мен төмен кідіріспен қамтамасыз етеді және секундына миллиондаған хабарламаны өңдей алады. Құрылғылар мен серверлер арасындағы байланыс үшін MQTT протоколы тиімді қолданылады, өйткені ол шектеулі ресурсы бар құрылғыларда жұмыс істеуге арналған.

Деректерді өңдеу қабаты. Apache Spark Streaming платформасы деректерді жадыда параллельді өңдеуге мүмкіндік береді, бұл нақты уақыт режимінде талдау жасауға қажет. Деректерді алдын ала өңдеу, шуылдан тазалау, нормализациялау және ерекшеліктерді шығару (feature engineering) осы кезеңде жүзеге асырылады. Бөлінген файлдық жүйе ретінде HDFS немесе бұлттық қоймалар (Amazon S3, Google Cloud Storage) қолданылады.

Талдау және визуализациялау қабаты. Өңделген деректер машиналық оқыту модельдерімен талданып, нәтижелер веб-интерфейс және мобильді қосымшалар арқылы пайдаланушыларға жеткізіледі. Grafana, Kibana және Apache Zeppelin сияқты құралдар көрнекі панельдерді (dashboard) жасауға мүмкіндік береді. Бұл шешім қабылдаушы мемлекеттік органдар мен халыққа ауа сапасы туралы өзекті ақпаратты жеделдетіп жеткізеді.

Атмосфералық ластану деңгейін болжау үшін машиналық оқыту мен терең оқыту әдістері қолданылды. Модельдерді бағалау кезінде уақыттық қатарлы деректер пайдаланылып, болжау дәлдігі детерминация коэффициенті (R^2) және орташа квадраттық қателік (MSE) сияқты көрсеткіштер арқылы өлшенді. Талдау барысында жеке базалық модельдер (сызықтық регрессия, Random Forest), терең оқыту архитектуралары (LSTM, CNN-RNN) және ансамбльдік алгоритмдер (Gradient Boosting, AdaBoost, XGBoost, LightGBM) салыстырылып, олардың ауа сапасы көрсеткіштерін болжаудағы тиімділігі бағаланды. Жүйенің жалпы құрылымдық моделі мен қабаттарының өзара байланысы (1-суретте берілген).



Сурет 1 - Үлкен деректерге негізделген нақты уақыттағы атмосфералық мониторинг жүйесінің архитектурасы

Үлкен деректерге негізделген нақты уақыттағы атмосфералық мониторинг жүйесінің архитектурасы - бұл атмосфералық параметрлерді (ауа сапасы, температура, ылғалдылық, қысым, ластаушы газдар, PM2.5/PM10 және т.б.) үздіксіз жинап, өңдеп, талдап және визуализация жасайтын күрделі көпқабатты жүйе.

Нәтижелер мен талқылау. Зерттеу әдістері ретінде ғылыми дереккөздерді талдау, заманауи сенсорлық және есептеу платформаларын салыстыру, сондай-ақ түрлі машиналық оқыту модельдерінің нәтижелерін жинақтап бағалау қолданылады.

Бұл зерттеудің нәтижелері экологиялық мониторинг жүйелерін жобалаумен айналысатын мемлекеттік органдар, ғылыми мекемелер және ақпараттық технология компаниялары үшін пайдалы болуы мүмкін, өйткені олар жүйенің техникалық құрылымын объективті бағалауға және нақты қалалық жағдайларға байланысты енгізу процесін бейімдеуге мүмкіндік береді.

Жүргізілген талдау атмосфералық ластануды болжауда машиналық оқыту мен терең оқыту әдістерінің жоғары тиімділік көрсететінін айқындады. Өртүрлі зерттеулерде алынған нәтижелер модельдің таңдалуы, кіріс деректердің сапасы мен болжанатын параметрдің сипатына байланысты дәлдіктің елеулі түрде өзгеретінін көрсетеді.

Ramadan және басқа авторлардың зерттеуінде өнеркәсіптік ортада сенсорлардан

жиналған деректер LSTM (Long Short-Term Memory), Random Forest және сызықтық регрессия модельдерімен өңделген. Нәтижесінде LSTM моделі температура мен ылғалдылықты болжауда $R^2 = 99\%$ мәніне жеткен, ал PM2.5 болжамында Random Forest моделі ең жоғары дәлдік көрсеткен ($R^2 = 84\%$). Бұл нәтиже метеорологиялық параметрлер мен бөлшектік ластану үшін әртүрлі модельдердің қажет болуын және IoT мен жасанды интеллектті біріктірудің практикалық тиімділігін көрсетеді.

Alrashidi және басқа авторлар жүргізген зерттеуде 12 мониторинг станциясынан жиналған уақыттық қатарлы деректерге сегіз түрлі регрессиялық машиналық оқыту моделі сыналған. Gradient Boosting және AdaBoost ансамбльдік алгоритмдері PM2.5 ауа сапасы индексі болжауда ең төмен квадраттық қателікке ($MSE < 1,5$) қол жеткізген [5]. Бұл нәтиже ансамбльдік модельдердің жеке базалық модельдерден гөрі тиімдірек екенін және бірнеше әлсіз болжаушыны біріктіру арқылы дәлдікті арттыруға болатынын дәлелдейді.

Терең оқыту бағытында конволюциялық нейрон желілері (CNN) мен рекуррентті нейрон желілерінің (RNN) гибриді архитектуралары ерекше орын алады. Utku және басқалардың зерттеуінде үш ел үшін (Үндістан, Италия, Германия) PM2.5 концентрациясын болжайтын CNN-RNN гибриді моделі жасалған. Бұл модель кеңістіктік ерекшеліктер мен уақыттық тәуелділіктерді бірге ескере отырып, дәстүрлі әдістермен салыстырғанда жоғары дәлдікке қол жеткізген [6]. Осылайша гибриді архитектуралар күрделі кеңістіктік-уақыттық деректерді өңдеуде айрықша тиімді болып табылады.

Алматы қаласының жағдайында Найзабаева және басқалардың зерттеуі айрықша назар аударады. Аталған жұмыста STL (Seasonal-Trend Decomposition using Loess) декомпозициясы, XGBoost және LightGBM модельдерінің ансамбльдік комбинациясы қолданылған. 20-дан астам мониторинг станциясынан алынған сағаттық деректер негізінде модель PM2.5 үшін $R^2 = 0,98$ -ден жоғары, ал CO үшін $R^2 = 0,83$ нәтижесін көрсеткен. Бұл нәтиже маусымдық және трендік құрауыштарды бөліп алу мен ансамбльдік алгоритмдерді біріктіру күрделі қалалық ортада болжау дәлдігін айтарлықтай арттыратынын дәлелдейді әрі әдістеме нақты уақыттағы ауа сапасын мониторингтеу мен ескерту жүйелерінде қолдануға жарамды деп бағаланған.

Алынған нәтижелерді салыстыру модель тиімділігінің бірнеше факторға тәуелді екенін көрсетеді. Біріншіден, болжанатын параметрдің сипаты маңызды: метеорологиялық көрсеткіштер (температура, ылғалдылық) тұрақты заңдылықтарға бағынатындықтан жоғары дәлдікпен болжанса, PM2.5 секілді бөлшектік ластану көздердің көптүрлілігіне байланысты болжау қиынырақ болады. Екіншіден, ансамбльдік әдістер мен декомпозиция тәсілдерін біріктіру жеке модельдермен салыстырғанда тұрақты түрде жоғары нәтиже береді. Үшіншіден, кіріс деректердің кеңістіктік және уақыттық ажыратымдылығы неғұрлым жоғары болса, болжам сапасы соғұрлым артады.

Сонымен қатар жүйенің практикалық тиімділігі тек модельдің дәлдігімен ғана емес, сондай-ақ деректерді жинау мен өңдеу инфрақұрылымының сапасымен анықталады. Apache Kafka мен Apache Spark Streaming платформаларын қолдану үлкен көлемдегі ағынды деректерді кідіріссіз өңдеуге мүмкіндік беретіндіктен, болжау нәтижелерін нақты уақыт режимінде алуға және халыққа жедел ескерту жасауға жол ашады. HDFS пен бұлттық қоймалар тарихи деректерді ұзақ мерзімді сақтауды қамтамасыз етіп, модельдерді қайта оқыту мен жетілдіруге негіз қалайды.

Талдау нәтижелері Қазақстан үшін атмосфералық ауаның сапасы өткір мәселе болып табылатынын растайды. 2021 жылы елдің жалпы шығарындысы 2,4 миллион тоннаны құраған, оның негізгі бөлігі Орталық өңірден шыққан. Ірі қалалардағы ауа сапасының деңгейі ДДҰ нормаларынан бірнеше есе асып түседі. Алматы қаласындағы зерттеулер ауа ластануының тыныс алу аурулары ($r = 0,97$) және өлім-жітім көрсеткіштерімен жоғары корреляцияда болатынын дәлелдеген. Бұл деректер интеллектуалды мониторинг жүйелерін енгізудің денсаулық сақтау тұрғысынан өзектілігін көрсетеді.

Осы жағдайда Big Data технологияларына негізделген нақты уақыттағы мониторинг жүйесін енгізу бірқатар артықшылықтарды ұсынады. Біріншіден, мемлекеттік станциялардың кеңістіктік шектеулерін арзан сенсорлар желісі арқылы өтеуге және қаланың әртүрлі аудандарын біркелкі қамтуға болады. Екіншіден, машиналық оқыту модельдері 1–24 сағат алға болжам жасауға мүмкіндік беріп, ластанудың шарықтау сәттерін алдын ала анықтайды. Үшіншіден, шешім қабылдау процесі автоматтандырылып, халықты ескерту жүйесі жұмыс істей алады. Төртіншіден, тарихи деректерді талдау негізінде ұзақ мерзімді экологиялық саясатты қалыптастыруға ғылыми негіз қаланады.

Дегенмен жүйені енгізуде бірқатар шектеулер мен мәселелерді ескеру қажет. Арзан сенсорлардың өлшеу дәлдігі калибрлеуді талап етеді, ал сенсорлық желінің тұрақты жұмыс істеуі үшін техникалық қызмет көрсету мен энергиямен қамту мәселелерін шешу керек. Сонымен қатар деректердің сапасы мен толықтығы модельдердің болжау дәлдігіне тікелей әсер ететіндіктен, деректерді алдын ала өңдеу мен валидациялау процестеріне ерекше көңіл бөлу қажет. Осы факторларды кешенді түрде ескеру жүйенің практикалық тиімділігін қамтамасыз етеді.

Қорытынды. Атмосфералық ластануды нақты уақыт режимінде мониторингтеу қазіргі заманғы экологиялық қауіпсіздікті қамтамасыз етудің басым бағыттарының бірі болып табылады. Қалалар мен өнеркәсіптік аймақтарда антропогендік факторлардың күшеюі атмосфералық ауаның сапасын үздіксіз бақылау қажеттілігін арттырады, бұл өз кезегінде жоғары өнімді есептеу және деректерді талдау технологияларын қолдануды талап етеді.

Жүргізілген талдау көрсеткендей, үлкен деректерді өңдеу технологияларын (Apache Kafka, Apache Spark, Hadoop Distributed File System – HDFS) сенсорлық желілермен және машиналық оқыту әдістерімен интеграциялау атмосфералық параметрлерді жоғары дәлдікпен өлшеуге, өңдеуге және талдауға мүмкіндік береді. Атап айтқанда, Apache Kafka нақты уақыттағы деректер ағындарын сенімді түрде тасымалдауды қамтамасыз етсе, Apache Spark және Flink осы деректерді ағындық және пакеттік режимде өңдеуге мүмкіндік береді. HDFS тәрізді таратылған сақтау жүйелері үлкен көлемдегі тарихи деректерді сақтауды және оларға масштабталатын қолжетімділікті қамтамасыз етеді.

Сонымен қатар, сенсорлық құрылғылардың кеңейтілген желісі (IoT-based environmental sensors) атмосфералық көрсеткіштерді кеңістіктік және уақыттық тұрғыдан жоғары жиілікте жинауға мүмкіндік береді. Бұл деректер машиналық оқыту алгоритмдерімен (Random Forest, Gradient Boosting, LSTM сияқты модельдер) өңделіп, ластану деңгейін болжау, трендтерді анықтау және аномалияларды ерте кезеңде диагностикалау міндеттерін шешеді.

Жүйенің ұсынылған төрт қабатты архитектурасы деректерді жинау, тасымалдау, сақтау, өңдеу және визуализациялау процестерінің толық циклін қамтиды. Мұндай құрылым модульдік сипатқа ие болғандықтан, жүйенің масштабталуын, икемділігін және сенімділігін арттырады. Әрбір қабат өзара үйлесімді жұмыс істей отырып, нақты уақыт режимінде жоғары өнімді мониторинг жүргізуге мүмкіндік береді.

Талдау нәтижелері LSTM, Random Forest, XGBoost, LightGBM және гибриді CNN-RNN модельдерінің ауа сапасы көрсеткіштерін болжаудағы жоғары тиімділігін растайды. Әсіресе ансамбльдік әдістер мен STL декомпозициясын біріктіру ($R^2 > 0,98$) күрделі қалалық орталарда, оның ішінде Алматы сияқты бассейндік жерлерде тиімді нәтиже көрсетеді [7].

Сонымен бірге болжау дәлдігінің болжанатын параметрдің сипатына, кіріс деректердің сапасына және модель архитектурасына тәуелді екендігі айқындалды, бұл нақты қолдану жағдайына сәйкес модельді таңдаудың маңыздылығын көрсетеді.

Қазақстан жағдайында мұндай интеллектуалды жүйелерді енгізудің практикалық маңыздылығы ерекше зор. Ірі қалалардағы ауа сапасының ДДҰ нормаларынан бірнеше есе асуы және ауа ластануының тыныс алу аурулары мен өлім-жітім көрсеткіштерімен жоғары корреляциясы ($r = 0,97$) интеллектуалды мониторинг жүйелерін денсаулық сақтау мен

экологиялық саясат тұрғысынан кезек күттірмейтін қажеттілікке айналдырады. Арзан сенсорлар желісі мемлекеттік станциялардың кеңістіктік шектеулерін өтеп, машиналық оқыту модельдері алдын ала ескерту жасауға және ұзақ мерзімді экологиялық шешімдер қабылдауға ғылыми негіз қалайды.

Жүйенің табысты жұмыс істеуі тек техникалық құралдарға ғана емес, сондай-ақ сенсорлық желіні дұрыс калибрлеуге, деректерді сапалы өңдеуге және болжау нәтижелерін шешім қабылдау процесіне тиімді интеграциялауға тәуелді болуында. Деректердің толықтығы мен сенімділігін қамтамасыз ету, инфрақұрылымды техникалық қолдау және модельдерді жүйелі түрде қайта оқыту жүйенің ұзақ мерзімді тиімділігінің кепілі болып табылады.

Болашақта зерттеулерді спутниктік деректерді (Sentinel-5P, MODIS), көлік ағыны мен өнеркәсіптік шығарындылар туралы ақпаратты қосу арқылы кеңейту, сондай-ақ түсіндірілетін жасанды интеллект (Explainable AI) әдістерін енгізу маңызды болып табылады. Түсіндірілетін модельдер болжамдардың негіздемесін айқындап, шешім қабылдаушы органдардың жүйеге деген сенімін арттырады. Мұндай жан-жақты әрі ғылыми негізделген жүйе қалалық ауа сапасын басқару үшін объективті шешімдер қабылдауға және халық денсаулығын сақтауға елеулі үлес қосады.

Алынған нәтижелер үлкен деректерді өңдеу технологияларын қолдану арқылы атмосфералық ластануды нақты уақыт режимінде мониторингтеу жүйесін құру, қазіргі экологиялық мәселелерді шешудегі ең тиімді цифрлық тәсілдердің бірі болып табылады.

Сондықтан мұндай жүйе IoT сенсорларынан, спутниктік деректерден және метеорологиялық станциялардан келетін үлкен көлемдегі гетерогенді деректерді үздіксіз жинап, оларды ағындық өңдеу технологиялары арқылы талдауға мүмкіндік береді.

Нақты уақыттағы деректерді өңдеу (мысалы, Apache Kafka, Apache Flink сияқты құралдар негізінде) атмосфералық жағдайдың өзгерісін бірден анықтап, қауіпті деңгейлерге жедел әрекет етуге жағдай жасайды. Ал машиналық оқыту алгоритмдері ластану деңгейін болжауға, оның таралу динамикасын анықтауға және ықтимал экологиялық тәуекелдерді алдын ала болжауға мүмкіндік береді.

Жүйенің басты артықшылығы - тек бақылаумен шектелмей, деректерге негізделген шешім қабылдауды қолдауында. Визуализация құралдары арқылы алынған ақпарат пайдаланушыларға, экологтарға және мемлекеттік органдарға түсінікті форматта ұсынылып, экологиялық саясатты тиімді жоспарлауға көмектеседі.

Сонымен қатар автоматты ескерту жүйелері ластану деңгейі шектен асқан жағдайда дер кезінде хабар беріп, адам денсаулығына және қоршаған ортаға келетін зиянды азайтуға ықпал етеді.

Қорытындылай келе, үлкен деректер мен нақты уақыттағы аналитиканы біріктіретін атмосфералық мониторинг жүйесі - бұл тек технологиялық шешім ғана емес, сонымен қатар қалалардың экологиялық қауіпсіздігін арттыратын, өмір сапасын жақсартуға бағытталған маңызды құрал. Болашақта мұндай жүйелер жасанды интеллектпен тереңірек интеграцияланып, толық автоматтандырылған «ақылды экологиялық басқару» деңгейіне жетуі мүмкін.

Алғыс

Авторлар қолжазбадағы барлық ақпарат үшін, соның ішінде плагиаттың болмауы, дәйексөздің дәлдігі және жасанды интеллект жасаған жалған фактілердің болмауы үшін жауап береді.

Әдебиеттер тізімі

1. Темирбеков Н. Оценка негативного воздействия загрязнения городского воздуха на здоровье населения с использованием метода машинного обучения / Н. Темирбеков, М. Темирбекова, Д. Тамабай, С. Касенов, С. Аскаров, З. Туkenова // Международный журнал

- экологических исследований и общественного здравоохранения. - 2023. - Том 20, № 18. - Статья 6770. - DOI: 10.3390/ijerph20186770.
2. Тасмурзаев Н. Недорогой IoT-датчик и предварительное технико-экономическое обоснование машинного обучения для мониторинга качества воздуха в салоне самолета: Пилотный пример из Алматы / Н. Тасмурзаев, Б. Амангельды, Г. Смагулова, З. Байгараева, А. Имаш // Датчики. - 2025. - Том 25, № 14. - Артикул 4521. - DOI: 10.3390/s25144521.
3. Гуйе А. Недорогая система мониторинга загрязнения окружающей среды в режиме реального времени на основе Интернета вещей с использованием ESP8266 NodeMCU / А. Гуйе, М.С. Дrame, С.А.А. Нианг // Измерение и контроль.-2024.-DOI: 10.1177/00202940241306690.
4. Рамадан М.Н.А. Система искусственного интеллекта на базе Интернета вещей для мониторинга и прогнозирования загрязнения воздуха в промышленной среде в режиме реального времени / М.Н.А. Рамадан, М.А.Х. Али, С.Ю. Ху, М. Альхедер, М. Альхербави // Экоотоксикология и экологическая безопасность. - 2024. - Том 283.- Статья 116856. - DOI: 10.1016/j.ecoenv.2024.116856.
5. Альрашиди Х. PM2.5: Прогнозирование индекса качества воздуха с использованием машинного обучения: данные станций мониторинга качества воздуха Кувейта / Х. Альрашиди, Ф.Н. Сибай, А. Абонама, М. Альрашиди, А. Альсабер // Устойчивое развитие. - 2025. - Том 17, № 20. - Статья 9136. - DOI: 10.3390/su17209136.
6. Утку А. Усовершенствованный мониторинг качества воздуха: гибридная модель CNN-RNN на основе глубокого обучения для прогнозирования PM2,5 / А. Утку, У. Кан, М. Альпсюлюн, Х.С. Балыкчи, А. Амузегар, А. Пилатин, А. Барут // Атмосфера. - 2025. — Том 16, № 9. - Статья 1003. - DOI: 10.3390/atmos16091003.
7. Найзабаева Л. Прогнозирование загрязнения воздуха в Алматы с использованием моделей машинного обучения Ensemble / Л. Найзабаева, Г. Сембина, А. Алиман, М. Сатымбеков, Н. Барлыкбай, Н. Сеилова // Журнал прикладных наук о данных. - 2025. - Том 6, № 4. - С. 2461-2476. - DOI: 10.47738/jads.v6i4.821.

References

1. Temirbekov N. Assessment of the Negative Impact of Urban Air Pollution on Population Health Using Machine Learning Method / N. Temirbekov, M. Temirbekova, D. Tamabay, S. Kasenov, S. Askarov, Z. Tukenova // International Journal of Environmental Research and Public Health. - 2023. - Vol. 20, No. 18. - Article 6770. - DOI: 10.3390/ijerph20186770.
2. Tasmurzayev N. A Low-Cost IoT Sensor and Preliminary Machine-Learning Feasibility Study for Monitoring In-Cabin Air Quality: A Pilot Case from Almaty / N. Tasmurzayev, B. Amangeldy, G. Smagulova, Z. Baigarayeva, A. Imash // Sensors. - 2025. - Vol. 25, No. 14. — Article 4521. - DOI: 10.3390/s25144521.
3. Gueye A. A low-cost IoT-based real-time pollution monitoring system using ESP8266 NodeMCU / A. Gueye, M.S. Drame, S.A.A. Niang // Measurement and Control. - 2024. - DOI: 10.1177/00202940241306690.
4. Ramadan M.N.A. Real-time IoT-powered AI system for monitoring and forecasting of air pollution in industrial environment / M.N.A. Ramadan, M.A.H. Ali, S.Y. Khoo, M. Alkhedher, M. Alherbawi // Ecotoxicology and Environmental Safety. — 2024. - Vol. 283. - Article 116856. - DOI: 10.1016/j.ecoenv.2024.116856.
5. Alrashidi H. PM2.5: Air Quality Index Prediction Using Machine Learning: Evidence from Kuwait's Air Quality Monitoring Stations / H. Alrashidi, F.N. Sibai, A. Abonamah, M. Alrashidi, A. Alsaber // Sustainability. - 2025. - Vol. 17, No. 20. - Article 9136. - DOI: 10.3390/su17209136.

6. Utku A. Advancing Air Quality Monitoring: Deep Learning-Based CNN-RNN Hybrid Model for PM2.5 Forecasting / A. Utku, U. Can, M. Alpsülün, H.C. Balıkçı, A. Amoozegar, A. Pilatin, A. Barut // Atmosphere. - 2025. - Vol. 16, No. 9. - Article 1003. - DOI: 10.3390/atmos16091003.
7. Naizabayeva L. Air Pollution Forecasting in Almaty using Ensemble Machine Learning Models / L. Naizabayeva, G. Sembina, A. Aliman, M. Satymbekov, N. Barlykbay, N. Seilova // Journal of Applied Data Sciences. - 2025. - Vol. 6, No. 4. - P. 2461–2476. - DOI: 10.47738/jads.v6i4.821.

П.А.Кожобекова*, М.А. Сыздыкова, О.М. Сулеймен, Ж.Д. Изтаев

к.т.н., доцент, pernesh-63@mail.ru, ЮКУ имени М.Ауэзова, Шымкент, Казахстан
магистрант, Ms.syzdykova@gmail.com, ЮКУ имени М.Ауэзова, Шымкент, Казахстан
магистр, преподаватель, jasiksuleimen42@gmail.com, ЮКУ имени М.Ауэзова, Шымкент,
Казахстан

к.п.н., ассоциированный профессор, zhalgasbek71@mail.ru, ЮКУ имени М.Ауэзова, Шымкент,
Казахстан

СОЗДАНИЕ СИСТЕМЫ МОНИТОРИНГА АТМОСФЕРНОГО ЗАГРЯЗНЕНИЯ В РЕАЛЬНОМ ВРЕМЕНИ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Аннотация

В статье рассматриваются вопросы создания системы мониторинга и прогнозирования загрязнения атмосферного воздуха в режиме реального времени с использованием технологий обработки больших данных (Big Data). Описывается архитектура системы, состоящая из этапов сбора, передачи, хранения, обработки и визуализации данных. Анализируется эффективность распределенных вычислительных платформ, таких как Apache Kafka и Apache Spark, при обработке экологических данных, полученных с помощью датчиков. Кроме того, возможности прогнозирования изменений качества воздуха с помощью алгоритмов машинного обучения демонстрируются на основе результатов LSTM, Random Forest, XGBoost, LightGBM и гибридных моделей CNN-RNN. В работе определено значение систем экологического мониторинга в контроле состояния окружающей среды и рассмотрена роль атмосферного загрязнения в снижении воздействия на здоровье человека. В условиях Казахстана обосновывается практическая значимость внедрения таких интеллектуальных систем в повышении экологической безопасности, оперативном анализе данных и принятии управленческих решений.

Ключевые слова: большие данные, загрязнение воздуха, мониторинг в реальном времени, машинное обучение, PM2.5, Apache Spark, Apache Kafka.

P.A.Kozhabekova*, M.A. Syzdykova, O.M. Suleimen, J.D. Iztayev

Candidate of Technical Sciences, Associate Professor, pernesh-63@mail.ru, M.Auezov SKU,
Shymkent, Kazakhstan

Master's student, Ms.syzdykova@gmail.com, M.Auezov SKU, Shymkent, Kazakhstan

Master, Lecturer, jasiksuleimen42@gmail.com, M. Auezov SKU, Kazakhstan

Candidate of Pedagogical Sciences, Associate Professor, zhalgasbek71@mail.ru, M.Auezov SKU,
Shymkent, Kazakhstan

DEVELOPMENT OF A REAL-TIME ATMOSPHERIC POLLUTION MONITORING SYSTEM USING BIG DATA PROCESSING TECHNOLOGIES

Abstract

The article discusses the issues of creating a system for monitoring and forecasting atmospheric air pollution in real time using big data processing technologies (Big Data). The architecture of the system is described, consisting of the stages of data collection, transmission, storage, processing and visualization. When

processing environmental data obtained using sensors, the effectiveness of distributed computing platforms such as Apache Kafka and Apache Spark is analyzed. In addition, the possibilities of predicting changes in air quality using machine learning algorithms are demonstrated based on the results of hybrid models LSTM, Random Forest, XGBoost, LightGBM and CNN-RNN. The paper defines the importance of environmental monitoring systems in controlling the state of the environment and examines the role of atmospheric pollution in reducing the impact on human health. The practical significance of the introduction of such intelligent systems in improving environmental safety in Kazakhstan, operational data analysis and management decision-making is substantiated.

Keywords: Big Data, atmospheric pollution, real-time monitoring, machine learning, PM2.5, Apache Spark, Apache Kafka.