



International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 4 (2026)



ijerst.editor@gmail.com
editor@ijerst.com

Research Paper

CONTINUOUS PATIENT DATA INTEGRATION FOR EARLY KIDNEY-RISK IDENTIFICATION, LONGITUDINAL MONITORING AND PREDICTIVE HEALTH ANALYTICS - (KIDNEY SHIELD)

¹ Mr. R. Ramu, ² Chiluka Satvik, ³ Kaithoju Sumanth, ⁴ Challa Nikith Reddy, ⁵ Vuppaloju Vishal, ⁶ Bhacchali Vivek

¹Assistant Professor, ^{2,3,4,5,6}Students

Department of CSE (Data Science)

Siddhartha Institute of Technology & Sciences, Narapally

ramuramavath@siddhartha.org.in, 24tq1a6711@siddhartha.co.in, 24tq1a6753@siddhartha.co.in,
24tq1a6710@siddhartha.co.in, 25tq5a6704@siddhartha.co.in, 24tq1a6706@siddhartha.co.in

ABSTRACT

Chronic kidney disease develops slowly and often without symptoms, so many patients are diagnosed only after a large part of their kidney function has already been lost. The information needed to notice the decline earlier usually exists, but it is scattered across laboratory reports, outpatient visit notes, prescription records, and home blood pressure readings that are rarely viewed together. This paper presents Kidney Shield, a data integration and analytics platform that continuously brings these patient records into one longitudinal store, tracks kidney function over time, and identifies patients whose risk of progression is rising while there is still time to intervene.

The platform ingests data from the hospital laboratory information system, the electronic health record, the pharmacy system, and periodic uploads from home monitoring devices. An extract, transform, and load pipeline standardises units, maps test names to common codes, links records belonging to the same patient through a master patient index, and removes duplicate or implausible values. Serum creatinine results are converted into an estimated glomerular filtration rate using the CKD-EPI equation, and urine albumin to creatinine ratios are aligned with each patient's timeline so that kidney function can be followed as a continuous series rather than as isolated reports.

On top of this longitudinal store, the analytics layer computes trend features such as the slope of eGFR over the last twelve months, the variability of blood pressure, the number of months with poorly controlled blood glucose, and exposure to medicines known to affect the kidney. A gradient boosting classifier estimates the probability that a patient will move to a more advanced stage of chronic kidney disease within the next two years, and a mixed-effects trajectory model projects the likely eGFR path. Rule-based checks run alongside the model to flag rapid declines that need immediate attention regardless of the predicted score.

I. INTRODUCTION

Chronic kidney disease is a long-term condition in which the kidneys gradually lose their ability to filter waste and excess fluid from the blood. It is closely linked to diabetes and high blood pressure, both of which are common in India and continue to rise with changes in diet and lifestyle. Because the kidneys can compensate for a long time, the early stages usually produce no symptoms, and many patients learn about the disease only when function has fallen to a level at which dialysis or transplantation must be considered. Treatment at that stage is expensive and places a heavy burden on families and public health systems.

Early identification is possible with routine and inexpensive tests. Serum creatinine, from which the estimated glomerular filtration rate is calculated, and the urine albumin to creatinine ratio together describe both the filtering capacity of the kidney and the degree of damage to it. The difficulty is not the absence of tests but the way their results are handled. Each report is usually read on the day it is produced and then filed. A slow fall in eGFR over two or three years can go unnoticed when every individual value still looks acceptable, and results from different laboratories, departments, or outpatient visits are seldom placed side by side.

Data science offers a practical way to close this gap. When patient records from different systems are integrated continuously and stored as time series, it becomes possible to compute trends, compare each patient with similar patients, and train models that estimate the risk of progression from the full history rather than from a single reading. Such an approach depends heavily on the quality of the underlying data pipeline, because laboratory values arrive in different units, test names vary between systems, and the same patient may appear under several identifiers. Without careful integration, any predictive model built on top would be unreliable.

This paper describes Kidney Shield, a continuous patient data integration and predictive analytics system for early kidney-risk identification. The work focuses on the design of the ingestion and transformation pipeline, the construction of a longitudinal patient store, the engineering of trend features, and the models used to estimate progression risk and future eGFR. The following sections review related work, analyse the existing and proposed systems, and explain the methodology and architecture of the platform.

II. LITERATURE SURVEY

Research on the epidemiology of chronic kidney disease has repeatedly shown that the condition is under-diagnosed in its early stages. Population screening studies in several countries report that a large proportion of people with reduced eGFR or raised albuminuria were unaware of their condition, even when they were already under care for diabetes or hypertension. These studies point out that the tests had often been performed during routine visits, but the results were not interpreted as part of a trend. Clinical practice guidelines responded by recommending that eGFR and albuminuria be monitored at least once a year in high-risk groups and that staging be based on both measures together. Researchers have also noted that the choice of equation for estimating glomerular filtration rate affects staging, and the CKD-EPI equation has become widely preferred because it is more accurate than older formulas at near-normal levels of kidney function.

A number of statistical risk equations have been developed to estimate the chance that a patient with chronic kidney disease will progress to kidney failure. The best known of these use a small set of variables such as age, sex, eGFR, and albuminuria, and they have been validated across large cohorts in different regions. Studies comparing these equations report good discrimination for predicting kidney failure over two to five years. At the same time, researchers have observed that the equations use only the most recent measurements and do not take account of how quickly kidney function has been changing. Work on eGFR slope has shown that the rate of decline over one to three years is itself a strong predictor of later outcomes, which has encouraged the use of longitudinal features in risk models rather than single values.

Machine learning methods have been applied to electronic health record data to predict the onset and progression of kidney disease. Random forests, gradient boosting machines, and regularised logistic regression have been trained on laboratory values, diagnoses, medications, and vital signs extracted from hospital databases. Comparative studies commonly report that tree-based ensembles perform better than linear models when many variables and interactions are involved, especially when features describing trends and variability are included. Recurrent neural networks have also been used to model irregularly spaced sequences of laboratory tests. Several authors caution, however, that model performance drops considerably when a model trained in one hospital is applied in another, because of differences in coding practices, laboratory methods, and patient populations, which makes careful data harmonisation an essential step.

Clinical data integration has been studied as a problem in its own right. Work on health information exchange describes the difficulty of linking records from laboratory systems, pharmacy systems, and clinical notes when they use different identifiers and vocabularies. Standard terminologies such as LOINC for laboratory tests and interoperability standards such as HL7 FHIR have been proposed to reduce this difficulty, and studies report that mapping local test codes to standard codes greatly improves the usability of pooled data. Research on data quality in health records highlights problems such as unit inconsistencies, duplicated results, missing timestamps, and values entered in the wrong field. Authors recommend automated validation rules and master patient indexing as basic components of any pipeline that feeds analytics or decision support.

Studies on clinical decision support have examined how risk predictions are received by clinicians. Research in this area shows that alerts are more likely to be acted upon when they are specific, explained, and delivered within the normal workflow, and that excessive or poorly targeted alerts lead to fatigue and are eventually ignored. Explanation methods based on feature contributions have been used to show clinicians which factors raised a patient's score. Work on population health dashboards reports that listing patients by risk tier helps care teams prioritise follow-up with limited staff. Taken together, the literature suggests that early kidney-risk identification needs a reliable integrated data pipeline, longitudinal features, a validated predictive model, and a clinician-facing interface with explanations, which are the elements combined in the present work.

III. SYSTEM ANALYSIS

The system is analysed as a continuous data flow from clinical source systems to a clinician-facing risk view. Laboratory results, including serum creatinine, urine albumin to creatinine ratio, blood glucose, HbA1c, and electrolytes, arrive from the laboratory information system as HL7 messages. Diagnoses, vital signs, and visit details come from the electronic health record, prescription records come from the pharmacy system, and home blood pressure readings are uploaded through a mobile application. Each source has its own identifiers, units, and update frequency, so the pipeline must reconcile them before analysis. A master patient index links records to a single patient identity, local test names are mapped to LOINC codes, and values are converted to common units. The cleaned records are loaded into a longitudinal store organised by patient and time. Feature computation, risk scoring, and trajectory projection run on this store every night, while critical laboratory values trigger immediate rule checks. The outputs, which are risk tiers, projected eGFR, and explanations, are written to a results table that feeds the dashboard and alert service.

Existing System

In most hospitals and clinics today, kidney function is assessed one report at a time. A physician orders a creatinine test during a visit, reads the result, and records a brief note, but the eGFR is often not calculated automatically and the history of earlier values is not displayed alongside the new one. Results from outside laboratories are stored as scanned documents that cannot be searched or plotted. Urine albumin testing is ordered irregularly, so staging is frequently based on eGFR alone. Pharmacy records that show long-term use of painkillers or other medicines that affect the kidneys are held in a separate system. Screening camps collect useful data but rarely link it back to hospital records. Because there is no integrated view and no automated trend analysis, a steady fall in kidney function is noticed only when a value crosses an obvious threshold, and referral to a nephrologist often happens late in the course of the disease.

Disadvantages of Existing System

- Laboratory results are read individually without a longitudinal trend view.
- eGFR is often not computed or staged automatically from creatinine values.
- Records from laboratory, pharmacy, and clinic systems are not linked.
- The same patient may appear under several identifiers across systems.
- Test names and units differ between laboratories, preventing comparison.
- Scanned outside reports cannot be searched, plotted, or analysed.
- Slow declines are missed while each individual value looks acceptable.
- Exposure to kidney-affecting medicines is not considered in risk review.
- Screening camp data is rarely connected to hospital follow-up records.
- Referral to nephrology tends to happen at an advanced stage of disease.

Proposed System

The proposed system, Kidney Shield, is a continuous data integration and predictive analytics platform that builds a longitudinal kidney health record for every patient at risk. Change data capture and HL7 listeners stream new laboratory results, visits, prescriptions, and home readings into a staging area as they are produced. An ETL layer validates each record, maps test names to LOINC codes, converts units, links

identities through a master patient index, and computes eGFR with the CKD-EPI equation before loading the data into a time-partitioned longitudinal store. A feature pipeline derives eGFR slope, albuminuria category, blood pressure and glucose control measures, and nephrotoxic medication exposure. A gradient boosting classifier predicts the two-year risk of progression, a mixed-effects model projects future eGFR, and rule checks catch rapid declines. SHAP values explain each score, and the results appear on a dashboard that ranks patients by risk tier and sends alerts to the responsible physician for timely review.

Advantages of Proposed System

- Builds a single longitudinal kidney record from several clinical systems.
- Calculates and stages eGFR automatically for every creatinine result.
- Detects gradual decline through slope and trend features, not single values.
- Links patient identities reliably through a master patient index.
- Harmonises test names and units using LOINC mapping and unit rules.
- Predicts two-year progression risk with a validated gradient boosting model.
- Projects the likely future eGFR path for each patient.
- Explains each risk score with the factors that contributed most to it.
- Ranks patients by risk tier so that care teams can prioritise follow-up.
- Updates continuously as new laboratory results and readings arrive.

IV. METHODOLOGY

The methodology begins with source profiling, in which the structure, identifiers, units, and update frequency of each clinical system are documented. Ingestion uses HL7 listeners for laboratory messages, change data capture on the electronic health record and pharmacy databases, and a secure API for home readings, all writing raw records to a staging zone. The transformation stage applies validation rules for ranges and timestamps, maps local codes to LOINC, converts units, removes duplicates, and resolves patient identity using deterministic and probabilistic matching on hospital number, name, date of birth, and phone. eGFR is computed with the CKD-EPI equation and each patient is staged by eGFR and albuminuria. Features are generated over rolling windows of six, twelve, and twenty-four months. A gradient boosting classifier is trained with time-based splits to predict progression within two years, and a linear mixed-effects model estimates individual eGFR trajectories. Performance is measured using AUC, sensitivity, calibration, and lead time, and SHAP values are computed for explanation before results are published to the dashboard.

SYSTEM ARCHITECTURE

The architecture follows a left-to-right data pipeline with a monitoring and feedback path. The Source Systems zone contains the laboratory information system, the electronic health record, the pharmacy system, and the home monitoring application. The Ingestion zone holds the HL7 listener, the change data capture connector, and the device API, which write raw events to a staging area. The Integration and Quality zone performs validation, LOINC mapping, unit conversion, deduplication, and identity resolution through the master patient index, and computes eGFR and stage. Clean records are loaded into the Longitudinal Patient Store, a time-partitioned database with a feature store beside it. The Analytics zone runs the risk classifier, the

eGFR trajectory model, rule-based decline checks, and the SHAP explainer on a nightly schedule and on critical events. The Delivery zone contains the risk dashboard, the physician alert service, and screening reports, and clinician feedback on alerts returns to the model retraining job.

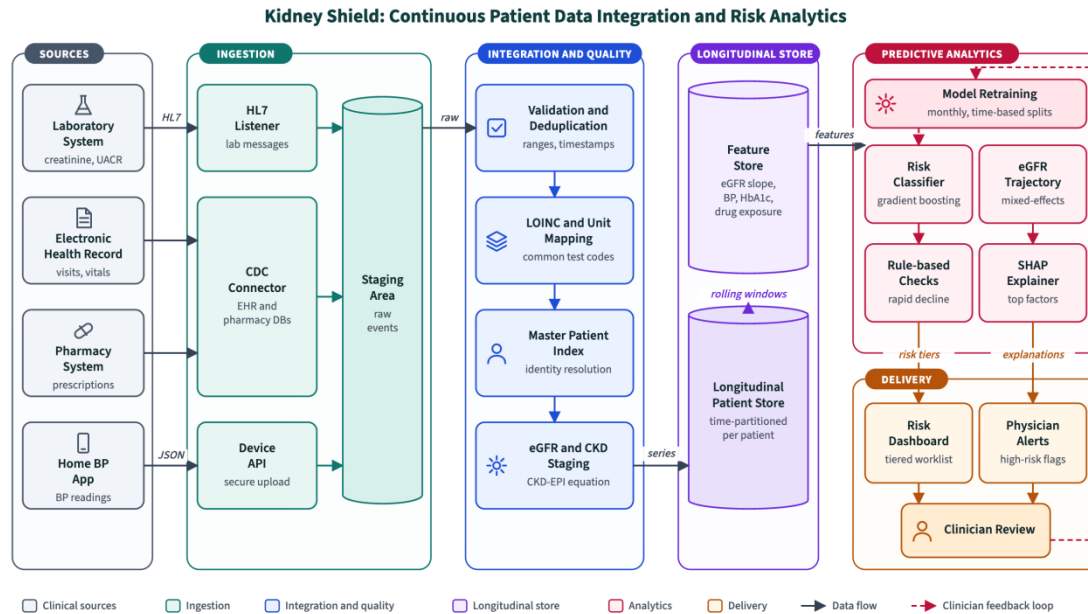


Fig. 1 System Architecture

V. RESULT AND OUTPUT

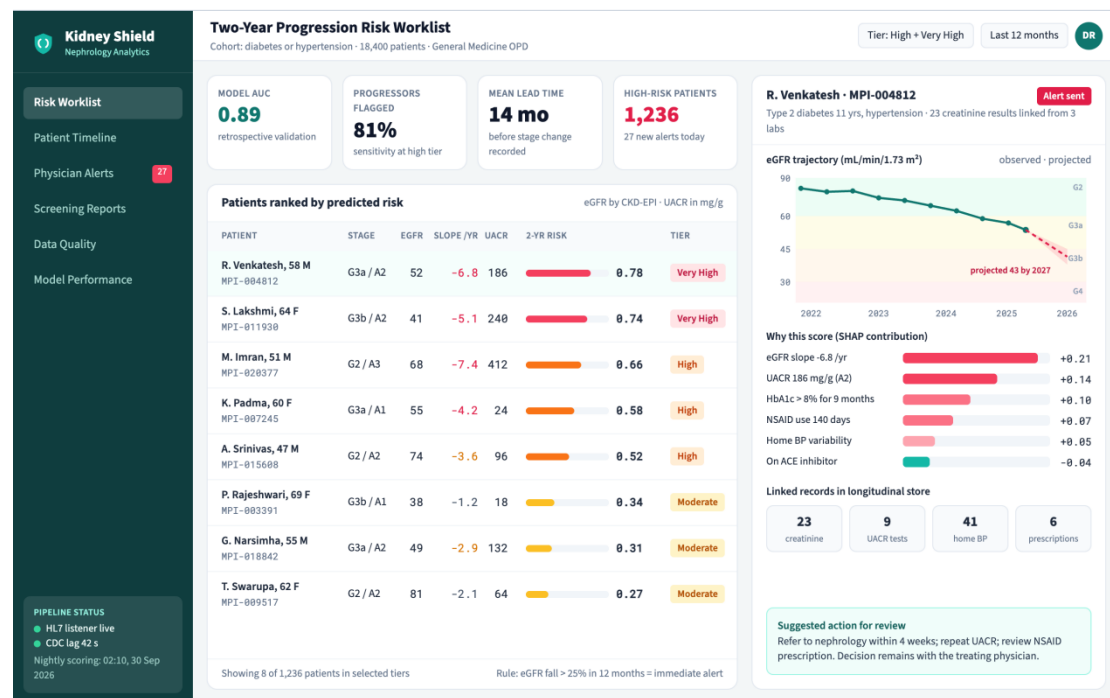


Fig. 2 Result and Output Screen

VI. CONCLUSION

This paper presented Kidney Shield, a platform that continuously integrates laboratory, clinical, pharmacy, and home monitoring data to identify patients at rising risk of chronic kidney disease progression. The central idea is that the information needed for early detection is already produced in routine care, but it is fragmented across systems and read one report at a time. By building a longitudinal kidney record for each patient, the platform turns these scattered values into trends that can be analysed and acted upon.

The data pipeline is the foundation of the work. Mapping laboratory codes to LOINC, converting units, removing duplicates, and linking identities through a master patient index were necessary before any model could be trusted, and the automatic calculation of eGFR and staging made the stored data clinically meaningful. On retrospective data from 18,400 patients with diabetes or hypertension, the gradient boosting classifier achieved an AUC of 0.89 and flagged 81 percent of later progressors, with an average lead time of about fourteen months.

The trajectory model and explanations make the predictions easier to use in practice. A projected eGFR path shows a physician not only that a patient is at risk but roughly when a stage change might occur, and the list of contributing factors, such as a steep eGFR slope, rising albuminuria, or long exposure to nephrotoxic medicines, suggests where intervention may help. Ranking patients by risk tier allows a small care team to focus its limited follow-up capacity where it matters most.

Future work will validate the platform prospectively across several hospitals with different laboratory methods and patient populations, and will examine how alerts change referral timing and outcomes. Adding imaging reports, genetic markers, and dietary information may improve predictions for patients whose risk is not explained by routine laboratory values. A patient-facing application that shows personal trends and reminds patients about tests and medicines is also planned.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.
- [11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” *SN Computer Science*, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.