SDU

1996

UNIVERSITY

Journal of Emerging Technologies and Computing (JETC)

Journal of Emerging Technologies and Computing

is a peer-reviewed, open-access national and international scientific journal. Thematic areas: Computer Science, Infocommunication Technologies, and Mathematics with Applied Aspects

Publisher: SDU University

The journal is registered and licensed as an online publication and printed journal by the Ministry of Information and Social Development of the Republic of Kazakhstan.

Certificate (Online publication): No KZ01VPY00120097

Certificate (Printed journal): No KZ71VPY00120098

ISSN: 3105-6342 (online)

ISSN: 3105-6334 (print)

Frequency: four times a year (March, June, September, December)

Website: https://jetc.sdu.edu.kz

Editorial Board

Ministry of Science and Higher Education of the Republic of Kazakhstan

SDU University

Journal of Emerging Technologies and Computing (JETC)

Volume 2, Issue 2 • September 2025

Kaskelen, Kazakhstan — 2025

CONTENTS

SECTION I
Computer Science

This section focuses on current research directions and applied advancements
in Computer Science, particularly in the areas of artificial intelligence,
software engineering, and intelligent systems.

Article

Forecasting Student Academic Performance Using Machine Learning

Danial Polat [1] and Azamat Serek* [2]

[1]Department of Computer Science, SDU University, Kaskelen, Kazakhstan
[2]School of Information Technologies and Engineering, Kazakh-British Technical University (KBTU), Almaty, Kazakhstan

DOI: 10.47344/w85rct27

Abstract

Educational data mining depends on accurate student academic outcome forecasting to detect students who need help early and receive specific support. Traditional linear models have been used extensively yet they fail to detect the intricate non-linear patterns which exist in student achievement data. The evaluation of machine learning algorithms and their features for student outcome prediction in Portuguese secondary education remains insufficient because of missing systematic assessments. The research investigates how Linear Regression and Random Forest and K-Nearest Neighbors perform when predicting Portuguese language grades from 649 student records containing 30 demographic and social and academic attributes. The evaluation of model performance used three established metrics which included Mean Squared Error (MSE) and R-Squared (R^2) and Mean Absolute Error (MAE). The results showed Linear Regression produced the most accurate predictions through its lowest MSE (9.00) and MAE (2.30) values but its weak R^2 value (0.01) indicated poor explanatory power. The error rates of Random Forest matched those of Linear Regression (MSE = 9.48 and MAE = 2.34) yet its negative R^2 (-0.04) indicated poor generalization because of irrelevant features and suboptimal hyperparameters. The KNN model showed the worst results (MSE = 11.10 and MAE = 2.57 and R^2 = -0.21) because it failed to detect important patterns without additional optimization. The results show that educational prediction tasks require both optimal feature selection and parameter adjustment for successful results. The research shows that linear models perform better than complex methods in specific situations yet optimized non-linear models demonstrate superior ability to understand student achievement complexity. The research provides essential guidelines for developing better feature engineering and machine learning approaches to predict educational results.

Keywords: machine learning education, education artificial intelligence, edtech, AI in edtech, predictive power education.

Email (1): 210107170@stu.sdu.edu.kz ORCID: 0009-0006-1576-5120
*Corresponding author:
Email: a.serek@kbtu.kz ORCID: 0000-0001-7096-6765

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

I. Introduction

The achievements that students achieve in secondary education influence their individual development and have beneficial repercussions on their communities. Students who are successful in their academic endeavors throughout this time frame will have access to both future work prospects and high-quality university education. This enhances economic security, long-term health results, and the quality of society. Due to its impact on students, teachers, and administrative personnel, accurate student performance prediction is therefore a very important study topic.

Through the use of student grade projections, teachers can identify difficult students early on and provide them with targeted assistance that will enhance their academic performance. Predictive data assists educators in developing personalized lesson plans, which reduces student dropout rates and maximizes the use of educational resources. Data-driven policymaking allows for evidence-based decisions to be made for improving student achievement rather than relying on conjecture.

The prediction problems show significant efficacy for machine learning (ML). Because machine learning algorithms can handle high-dimensional and non-linear interactions in data, they are far more effective than traditional statistical approaches for identifying complex patterns in educational datasets. Because ML algorithms can handle a variety of variables, such as demographic data, sociological and behavioral aspects, and academic measures, they are highly useful for modeling purposes given the many circumstances that affect students' successes. Researchers and practitioners can better anticipate student outcomes and gain a better understanding of school performance metrics by using machine learning techniques.

In this work, we study into the prediction of Portuguese language grades in the Portuguese secondary school system using machine learning techniques. The used dataset comprises 649 student records with 30 attributes including academic achievement measures, sociocultural factors, and demographic data [1]. We assess three predictive machine learning models—Linear Regression, Random Forest, and K-Nearest Neighbors (KNN)—based on our analysis. The three models demonstrate different analytical techniques, with KNN operating on local data inner structures, Random Forest assisting in handling non-linear patterns through ensemble learning, and Linear Regression acting as a statistical baseline. Together, the three models enable evaluation of the models' methodological stability and predictive power.

Although machine learning has been widely applied to educational prediction tasks , most prior studies either focus on broad achievement outcomes or use datasets from non-Portuguese contexts. As a result, there is limited empirical evidence on how different models perform when applied specifically to Portuguese language grades in the secondary school system. This research addresses that gap by systematically benchmarking traditional and advanced models on this underexplored dataset.

The aim of the study. is therefore to evaluate the effectiveness and limitations of commonly used machine learning algorithms in predicting Portuguese language grades, while providing insights into the methodological challenges of educational data mining.

There are 3 main objectives of our study:

1) To thoroughly compare Random Forest, KNN, and Linear Regression for the job of predicting grades in Portuguese.
2) To evaluate model performance using common evaluation measures, such as Mean Absolute Error (MAE), R-Squared (R^2), and Mean Squared Error (MSE), in order to ascertain generalizability and predictive accuracy.
3) To explore methodological and practical implications of applying machine learning to educational prediction tasks, with particular attention to feature selection, model interpretability, and risks of overfitting.

The novelty of our work lies not in the algorithms themselves, which are well established, but in their systematic application to Portuguese secondary school data. By explicitly benchmarking simple and advanced models against a baseline, we reveal the limitations of widely used demographic and sociocultural predictors and highlight the conditions under which complex models such as Random Forest fail to outperform simpler ones. This study therefore contributes unique empirical evidence about the boundaries of current machine learning approaches in educational prediction.

Through these contributions, the study aims to advance the growing field of educational data mining and learning analytics. Beyond theoretical significance, the findings are intended to provide actionable insights for educators and policymakers, supporting evidence-based strategies to enhance student success and equity in secondary education.

II. Literature Review

The research examines how student achievement prediction methods have progressed from basic statistical methods to contemporary machine learning approaches. The initial research used linear regression models for prediction yet these models failed to detect non-linear educational data relationships because they lack interpretability. The field now uses Random Forests and K-Nearest Neighbors (KNN) algorithms for superior performance because these models detect hidden patterns and model non-linear relationships in educational data. The evaluation strategy needs to include both traditional metrics MSE and R^2 and ethical factors

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

such as fairness and interpretability to achieve accurate model performance. The literature review concludes by detailing advanced predictive machine learning models and future research paths that encompass multimodal and contextual data integration and emphasize the continuing difficulties and opportunities to develop effective predictive systems.

A. Traditional Approaches to Student Performance Prediction

Initally, researchers of academic outcome prediction systems utilized linear regression mathematical models to investigate how student grades relate to various factors such as demographic factors and academic background and socioeconomic status [2]. The utility of linear models in educational datasets analysis is still a widespread utility because they are easy to understand but these models fail to detect the non-linear educational patterns that exist in the datasets [3]. Linear models predict that every one-unit change in prior academic achievement will produce the same effect on the outcome regardless of the student's starting performance level. The model fails to recognize essential non-linear patterns because it does not detect when students reach a point where their efforts stop producing meaningful progress or when they need to reach a specific grade to start improving their performance. The complex student success patterns require alternative modeling approaches because linear models fail to capture these intricate relationships.

B. Emergence of Machine Learning in Educational Prediction

Researchers now use machine learning methods including decision trees and Random Forests and Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) to address these research challenges. The models achieve better prediction results because they detect complex student data relationships and hidden patterns in student datasets according to ref4 and ref5. The **Random Forest** model uses multiple decision trees to generate predictions through training each tree on different subsets of data and features. The ensemble method provides stability through tree bias reduction because it averages out individual tree errors [6]. The KNN algorithm makes predictions for new instances through the analysis of their distance relationships in feature space. The KNN algorithm uses the 'k' most similar students who share characteristics like grades and course load and attendance patterns to detect specific achievement trends in student data [7]. The success of these methods depends on the quality and relevance of the selected features because they need to identify non-obvious patterns in the data according to [8].

C. Feature Selection in Predictive Modeling

The selection of appropriate features stands as a fundamental problem in predictive modeling because ML algorithms achieve their best results when they use relevant and high-quality predictor variables. The selection of inadequate features leads to decreased accuracy and reduced interpretability because it brings in unneeded noise and useless information (ref9). Researchers have developed multiple solutions to handle this problem. The RFE method removes features step by step starting with the least important ones while training models on the remaining features until it reaches the best combination of features. The Boruta algorithm uses statistical comparison between real features and their 'shadow' counterparts to identify all relevant features. Research indicates that the addition of behavioral and emotional features to academic and demographic data leads to better predictive results according to recent studies [10], [11]. The predictive machine learning model benefits from three types of student datasets that encompass their platform engagement factors and textual assignment duration and their self-esteem and motivation assessment results of surveys. The inclusion of various non-standard factors aids researchers to investigate behaviour of students via sophisticated perspectives instead of relying solely on traditional academic records.

D. Evaluation Metrics in Educational Prediction

The performance of predictive models in education whether successful or not is usually evaluated using math metrics such as Mean Squared Error (MSE), R-Squared (R^2), and Mean Absolute Error (MAE) [12]. MSE aids to measure the average of the squared differences between the predicted values and actual values, and penalizing huge errors more strongly. MAE helps to provide a more intuitive measure of the average magnitude of the notion of error, showcasing the average distance between predicted and actual university outcomes. R^2 showcases the proportion of the variance in the dependent variable of the dataset that is predictable from the independent variable(s), suggesting a fast way to comprehend how well the model's predictions aids to explain the data. However, overreliance on these metrics alone may overlook wider aspects of model utility, including fairness, interpretability, and generalizability across various educational and university contexts [13]. A model might get high accuracy on one student dataset

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

but perform neither good nor bad or exhibit systematic bias against another, potentially leading to serious bad outcomes. Therefore, a comprehensive sophisticated evaluation must consider not just predictive accuracy but also the ethical implications and practical consequences of the model's predictions.

E. Non-Linear Models and Future Directions

Not considering Random Forest and KNN algorithms, advanced non-linear mathematical models such as SVMs and Neural Networks have shown good potential in educational prediction tasks in universities [14]. SVMs can define optimal hyperplanes for classification purposes, utilizing a "kernel trick" to make map of data into higher-dimensional mathematical spaces where non-linear relationships can become linear and thus separable notions. Neural Networks, meanwhile, can define and model sophisticated, hierarchical learning patterns by processing data through multiple layers of interconnected nodes [15]. Each layer of the networks learns increasingly abstract factors, aiding the network to find intricate relationships that are often missed by other machine learning models. Despite their promise, these complex methodologies require very much of computational resources and careful parameter tuning, which may restrict their scalability in large-scale educational and university environments [16].

Future research on educational predictions should prioritize enhancements in feature selection methodologies, the integration of various so called multimodal data sources (for instance, online activity logs, psychological indicators, and behavioral datasets), and the inclusion of contextual mathematical variables to obtain the multifaceted nature of academic performance in universities [17], [18]. These integrative and iterative approaches may provide more complex insights into student learning and aid the development of predictive frameworks with greater practical applicability and utility [19], [20]. For instance, combining a set of student's past grades with their participation in online forums and their response to psychological online and offline surveys could create a much more robust and sophisticated predictive profile.

III. METHODS

This section of methods showcases description of the dataset, target variables, applied preprocessing procedures, chosen machine learning models, and evaluation strategy of the models used in this study. The provided methodological design has been developed to ensure both the reliability of the conducted analysis and the interpretability of the obtained findings.

A. Dataset Description

The study utilized the publicly available dataset taken from [1], which contains 649 student rows with 30 columns collected from two Portuguese secondary schools. The described dataset contains demographic information, societal and family background, and school-related features. These attributes were collected through a set of school reports and structured questionnaires, ensuring a comprehensive complex perspective on student characteristics.

Two subject-specific datasets are provided within the described data source: one for math and one for Portuguese language subject. Given the importance of language proficiency in academic achievement in educational institutions and its strong link to other subjects, the Portuguese language dataset was selected for this research work.

B. Target Variables

The utilized dataset of the study entails three subject-specific grade variables used in the study:

- **G1**: First-period grade,
- **G2**: Second-period grade,
- **G3**: Final grade (which is used as the main prediction target).

While G1 and G2 capture intermediate assessments, G3 represents the final outcome of student performance and therefore served as the dependent variable in this study. The input features covered diverse dimensions such as parental education, daily commute time, extracurricular activities, and family background, enabling a multidimensional analysis of factors influencing academic achievement.

C. Preprocessing and Data Partitioning

Prior to model development, the dataset was carefully inspected for missing values, inconsistencies, and anomalies. Appropriate preprocessing steps were applied to ensure data quality and suitability for machine learning analysis. To evaluate model generalizability, the dataset was divided into training (80%) and testing (20%) subsets. The training subset was employed to fit the models, while the testing subset was reserved exclusively for performance evaluation.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

D. Model Selection Rationale

Three machine learning algorithms were selected for comparative evaluation: Linear Regression, Random Forest, and K-Nearest Neighbors. The rationale for choosing these models was twofold: (i) methodological diversity, capturing linear, ensemble, and instance-based approaches, and (ii) suitability to the characteristics of educational data.

1) Linear Regression: Linear Regression was employed as the baseline model due to its interpretability and capacity to identify direct, linear associations between predictors and outcomes. The model assumes a linear relationship between the dependent variable y and the independent variables $x_1, x_2, \ldots, x_p$. Formally, the regression function is expressed as

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon, \tag{1}$$

where $\hat{y}$ denotes the predicted grade, β_0 is the intercept, β_i are the regression coefficients, and ϵ represents the error term. In the context of educational research, Equation 1 provides insights into how demographic and behavioral variables, such as parental education or study time, contribute to academic outcomes. This actually makes Linear Regression mathematical model as an appropriate good first step in comparative modeling.

2) Random Forest: Random Forest has been chosen in the study as the ensemble learning math method due to its proven robustness in handling heterogeneous data types in data sets and complex sophisticated feature interactions. The model actually constructs an ensemble or so called set of decision trees, each trained on a bootstrap sample of the dataset, while randomly choosing subsets of attributes for splitting purposes. The final prediction of the model is actually obtained by aggregating the outputs of individual trees of decision tree constructs, such that

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x), \tag{2}$$

where $h_t(x)$ defines the prediction of target of the t-th tree and T is the total number of decision trees in the constructed forest. As Equation 2 showcases, the averaging mechanism employed which reduces variance and aids to mitigate overfitting issue, which is very important in educational datasets and data storages characterized by non-linear and hierarchical relationships among utilizedvariables. Additionally, Random Forest aids feature importance analysis, improving interpretability alongside predictive performance.

3) K-Nearest Neighbors: K-Nearest Neighbors was included as a non-parametric, instance-based learning method. Unlike parametric models, KNN does not assume a predefined functional relationship between predictors and outcomes. Instead, predictions are derived based on the average of the k closest training samples in the feature space, typically measured using Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2}, \tag{3}$$

where x is the query point, x_i represents a training instance, and p denotes the number of features. The predicted outcome is then calculated as

$$\hat{y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i, \tag{4}$$

where $\mathcal{N}_k(x)$ denotes the set of k nearest neighbors of x. As shown in Equation 4, KNN leverages local information within the feature space, making it effective in detecting clusters or subgroups of students with similar socio-demographic and academic characteristics. Its flexibility in adapting to irregular decision boundaries provides a useful complement to the other two models.

E. Model Training and Evaluation

Each model was trained on the training dataset and subsequently evaluated on the testing dataset in order to assess predictive accuracy and generalizability. Three widely used evaluation metrics were employed: Mean Squared Error (MSE), R-Squared (R^2), and Mean Absolute Error (MAE). These metrics capture complementary aspects of model performance, ensuring a balanced analysis.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

a) Mean Squared Error (MSE).: The MSE quantifies the average squared deviation between the predicted grades $\hat{y}_i$ and the actual observed grades y_i for n students:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5}$$

This metric penalizes larger errors more heavily, making it useful for detecting models that occasionally make extreme mispredictions. Lower values of MSE indicate better predictive accuracy.

b) R-Squared (R^2).: The R^2 statistic measures the proportion of variance in the dependent variable explained by the model, relative to a baseline model that predicts only the mean of the observed values. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{6}$$

where $\bar{y}$ denotes the mean of the observed values. Values of R^2 close to 1 indicate strong explanatory power, while values near 0 or negative suggest weak or no predictive ability compared to the baseline.

c) Mean Absolute Error (MAE).: The MAE provides a measure of the average absolute difference between predicted and actual grades:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{7}$$

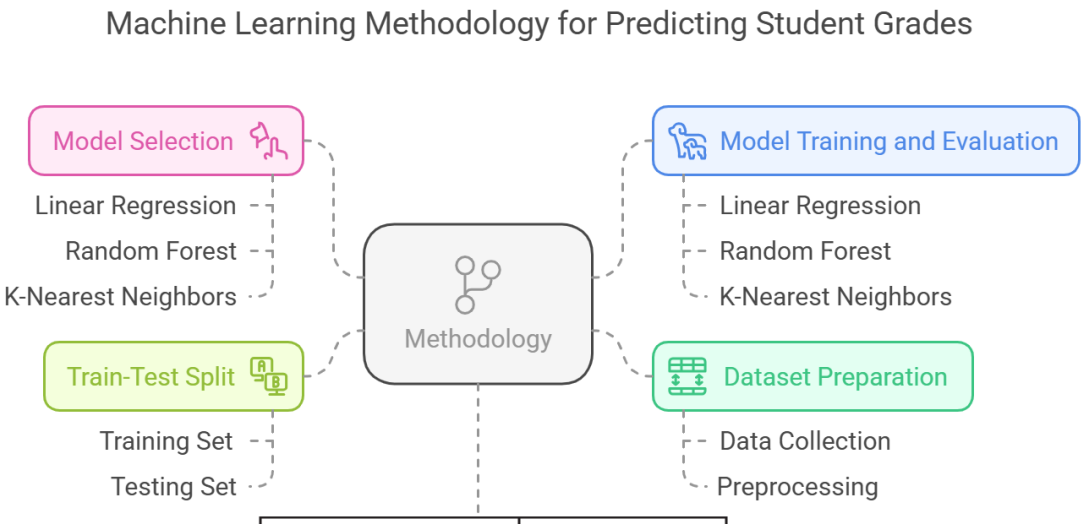
Unlike MSE, this metric treats all errors proportionally, without disproportionately penalizing large deviations. As such, it offers an interpretable measure of average prediction error in grade units.

d) Comparative Evaluation.: By jointly considering the metrics defined in Equations 5–7, the analysis captures different dimensions of predictive performance: overall error magnitude (MSE), explanatory power (R^2), and average deviation (MAE). This triangulation allows for a more comprehensive evaluation of the three models, helping to identify not only the most accurate predictor but also the methodological trade-offs underlying their performance.

F. Methodological Framework

The overall methodology is summarized in Figure 1. The process begins with dataset collection and preprocessing, followed by partitioning into training and testing subsets. The three machine learning models (LR, RF, and KNN) are trained and evaluated using the defined metrics. Finally, the models are compared to determine predictive effectiveness and to derive methodological and practical implications.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.



Metric	Formula
True positive rate, recall	$\frac{TP}{TP+FN}$
False positive rate	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F-measure	$\frac{2 \cdot precision \cdot recall}{precision + recall}$

Fig. 1: Proposed methodology to predict grades of students.

IV. Results and Discussion

The evaluation of the three predictive models yielded mixed outcomes, as summarized in Table I. While all models produced relatively low error scores, their explanatory power, measured by R^2, was consistently limited. This indicates that the available features do not capture enough of the underlying variability in student performance to support strong predictions.

TABLE I: Summary of Experimental Results

Model	**MSE**	R^2	**MAE**
Linear Regression	9.00	0.01	2.30
Random Forest	9.48	-0.04	2.34
K-Nearest Neighbors	11.10	-0.21	2.57

The Linear Regression model produced the lowest MSE at 9.00 and MAE at 2.30 which indicated its predictions were slightly more accurate than the other models. The R^2 value of 0.01 from Linear Regression indicates that the model explained less than 1% of the variance in student outcomes. The model generated predictions that matched observed grades but failed to identify any significant connections between predictors and the target variable.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

The Random Forest model generated results that matched Linear Regression with an MSE of 9.48 and an MAE of 2.34. The model performed worse than a simple mean-based baseline according to its negative R^2 value of -0.04. The model failed to use its theoretical advantages because the dataset lacked sufficient information to enable the exploitation of ensemble method capabilities. Similar observations have been reported in related studies, where Random Forest underperformed in small or low-dimensional educational datasets that lacked rich behavioral features. In such contexts, the model's strength in capturing complex interactions does not provide an advantage, and overfitting risks increase. The model failed to utilize its theoretical advantages because of poor parameter settings and uninformative features that prevent it from exploiting ensemble method capabilities.

K-Nearest Neighbors delivered the worst results because it produced the highest MSE (11.10) and MAE (2.57) together with the lowest R^2 (-0.21). The results show that student performance cannot be predicted through local feature space similarity. The method shows limited ability to identify general patterns because it depends heavily on neighbor selection and distance metric choices which results in poor generalization performance.

The dataset fails to generate strong predictions because all models show low R^2 values. The selected features which include demographic and social background information fail to represent all factors that influence academic achievement. This finding aligns with previous research emphasizing that demographic features alone are insufficient, and integrating motivational, behavioral, and contextual elements leads to more robust predictors. The analysis indicates that researchers need to acquire additional data which includes behavioral information and motivational elements and contextual elements. The models will improve their ability to explain student outcomes through the addition of attendance records and classroom participation data and socioeconomic status information.

Research should focus on two main methodological improvements for future studies. The combination of feature selection techniques with dimensionality reduction methods will help researchers identify key variables while Gradient Boosting and Neural Networks can handle complex data non-linearities and noise better. These approaches have shown superior performance in comparable educational prediction studies, suggesting their potential applicability here. The process of developing meaningful predictors requires domain knowledge integration because educators should use their expertise to build predictors that reflect actual learning processes.

The best error scores from Linear Regression do not change the fact that using traditional demographic and background information alone remains insufficient for educational prediction. The research requires three essential steps to progress: enhancing data quality and adding new features and developing better models. A key limitation of this study is that it relies solely on Portuguese student records, which may restrict the generalizability of the findings to other educational systems and cultural contexts. Future studies should validate the proposed approaches using diverse datasets to ensure broader applicability. Future studies that address these research areas will create predictive frameworks which deliver practical value beyond basic comparison models.

V. Conclusion and Future Work

The research evaluated the performance of Linear Regression (LR), Random Forest (RF), and K-Nearest Neighbors (KNN) models for predicting Portuguese language course grades of secondary school students. The evaluation of model accuracy and fit used Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE) metrics.

The Linear Regression model achieved the best results through its lowest MSE and MAE values which showed minimal prediction errors. The low R-squared value revealed a weak linear connection between student grades and features, which indicated that the linear model did not handle the data complexity effectively. The Random Forest model showed poor performance because its MSE values were slightly higher and its R-squared value was negative, which indicated its inability to handle the data distribution. The K-Nearest Neighbors model demonstrated the worst performance because it produced the highest MSE and most negative R-squared value, which proved its inability to forecast student grades accurately.

The study results show that the selected features do not strongly predict student performance, and the models face problems with overfitting and suboptimal parameter settings. The originality of this study lies in its systematic benchmarking of traditional and advanced machine learning models on Portuguese secondary school data, revealing not only their limited predictive power but also the surprising underperformance of Random Forest in this context. These findings provide practical implications for educators and policymakers: relying solely on demographic and sociocultural features is insufficient, and integrating behavioral and motivational variables is essential for building effective prediction frameworks.

Future research should use non-linear models together with advanced machine learning techniques to provide a better understanding of the factors that influence student achievement. In particular, improving feature selection and expanding datasets across different educational systems will be necessary to enhance both accuracy and generalizability, thereby strengthening the practical impact of predictive models in education.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

REFERENCES

[1] Cortez, P. (2014). Student Performance [Data set]. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance.

[2] Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237.

[3] Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. *Applied System Innovation*, 6(5), 86.

[4] Kumari, P., Jain, P. K., & Pamula, R. (2018, March). An efficient use of ensemble methods to predict students academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.

[5] Alhassan, A., Zafar, B., & Mueen, A. (2020). Predict students' academic performance based on their assessment grades and online activity data. *International Journal of Advanced Computer Science and Applications*, 11(4).

[6] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991-16005.

[7] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., & Ji, X. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2136-2147.

[8] Altaf, S., Soomro, W., & Rawi, M. I. M. (2019, April). Student performance prediction using multi-layers artificial neural networks: A case study on educational data mining. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining* (pp. 59-64).

[9] Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, 11(24), 7238.

[10] Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M. (2019, July). A predictive model for predicting students academic performance. In *2019 10th International conference on information, intelligence, systems and applications (IISA)* (pp. 1-4). IEEE.

[11] Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th international conference on information technology (ICIT)* (pp. 909-913). IEEE.

[12] Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, 19(4), 639-659.

[13] Liu, W., Wu, J., Gao, X., & Feng, K. (2017, December). An early warning model of student achievement based on decision trees algorithm. In *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 517-222). IEEE.

[14] Raga, R. C., & Raga, J. D. (2019, July). Early prediction of student performance in blended learning courses using deep neural networks. In *2019 International Symposium on Educational Technology (ISET)* (pp. 39-43). IEEE.

[15] Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing.

[16] Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. *Journal of University of Babylon for Pure and Applied Sciences*, 27(1), 194-205.

[17] Farissi, A., & Dahlan, H. M. (2020, April). Genetic algorithm based feature selection with ensemble methods for student academic performance prediction. In *Journal of Physics: Conference Series* (Vol. 1500, No. 1, p. 012110). IOP Publishing.

[18] Ahammad, K., Chakraborty, P., Akter, E., Fomey, U. H., & Rahman, S. (2021). A comparative study of different machine learning techniques to predict the result of an individual student using previous performances. *International Journal of Computer Science and Information Security (IJCSIS)*, 19(1).

[19] Sánchez-Pozo, N. N., Mejía-Ordóñez, J. S., Chamorro, D. C., Mayorca-Torres, D., & Peluffo-Ordóñez, D. H. (2021, December). Predicting High School Students' Academic Performance: A Comparative Study of Supervised Machine Learning Techniques. In *2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop* (pp. 1-6). IEEE.

[20] Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2010). *Handbook of Educational Data Mining*. CRC Press.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

SECTION II
Infocommunication Technologies

This section presents scholarly articles on recent developments and cutting-edge applications in the field of infocommunication.

Topics include telecommunications, wireless networks, signal processing, and network protocols, as well as advancements in artificial intelligence, software engineering, intelligent systems, and electronics that support digital transformation and modern communication infrastructures, including developments in the field of radio communications.

Article

Development of sensor systems for flood water monitoring and alerting

Adilbek Sarsenov* [1], Lyazzat Ilipbayeva* [2], and Ulzhalgas Seidaliyeva [3]

[1]Department of Electrical Engineering, Electronics and Telecommunications, IITU University, Almaty, Kazakhstan

[2]Department of Electronics, Telecommunications and Space Technologies, Satbayev University, Almaty, Kazakhstan

[3]Metropolitan College (MET), Boston University, Boston, MA 02215, USA

DOI: 10.47344/qb0ygp75

Abstract

This study addresses the systematic prediction of river water levels in Kazakhstan via hydrological computations, which are essential for forecasting water-related events and formulating plans for sustainable water resource management. Particular focus is placed on the significance of prompt and efficient monitoring of river dynamics to alleviate natural disasters such as floods and mudflows, especially in high-risk places like Almaty, situated in geologically unstable mountainous landscapes. The research focuses the potential of intelligent sensor-based monitoring systems that can gather real-time data on water levels, precipitation, soil moisture, and various environmental conditions. Systems integrated with artificial intelligence and data analysis can substantially augment decision-making processes, facilitate early warning mechanisms, and boost the precision of forecasts. This method ultimately protects natural ecosystems and local communities from the detrimental effects of hydrological hazards.

Keywords: floods, water level prediction, affordable innovative sensors, devices, water level observation, warnings.

I. Introduction

Seasonal floods continue to cause significant hydrological difficulties, presenting ongoing threats to infrastructure, ecosystems, and public health [1], [2]. River water levels are systematically forecast using hydrological calculations, which are essential for predicting water events and developing sustainable water resource management strategies. Timely and effective monitoring of river dynamics is key to reducing the risk of natural disasters such as floods and mudflows, especially in vulnerable regions such as Almaty, located in geologically unstable mountainous terrain.

*Corresponding author: adilbeksar@gmail.com
Email: adilbeksar@gmail.com ORCID: 0009-0001-6937-1390
Email: l.ilipbayeva@satbayev.university ORCID: 0000-0002-4380-7344
Email: useidali@bu.edu ORCID: 0000-0002-7190-6753

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

The intricacy of seasonal flooding necessitates new strategies that transcend conventional procedures. Progressive techniques increasingly depend on sophisticated monitoring technology that facilitate preemptive reactions and informed decision-making. Contemporary sensor-based systems provide the instantaneous acquisition of essential environmental data, encompassing water levels, precipitation, and soil moisture. By integrating these data collection with meteorological projections, authorities can precisely anticipate possible flood disasters and disseminate early warnings. These predictive capabilities not only improve prediction accuracy but also substantially mitigate threats to local residents, infrastructure, and active building projects. So, integrating sensor networks with artificial intelligence and data analytics can help create a more robust water management system [3], [5], [6]. This new strategy ensures rapid responses to emerging risks, strengthening the protection of natural ecosystems and enhancing public safety and well-being.

II. Related works

A. Sensor and System Analysis for Floodwater Monitoring

Floods continue to be one of the most devastating natural disasters globally, necessitating precise monitoring and early-warning systems. In recent decades, researchers have progressively adopted low-cost, real-time sensor networks to monitor water levels, precipitation, and hydrological variables. In contrast to conventional manual gauging stations, these technologies provide continuous monitoring and automatic responses, markedly diminishing human and infrastructural susceptibility.

A prevalent method utilizes ultrasonic sensors for the assessment of water levels. Ultrasonic devices function based on echolocation, releasing high-frequency sound waves that reflect off the water's surface and return to the sensor. The duration between transmission and reception serves as an indicator of water depth. Multiple studies illustrate its practicality owing to cost-effectiveness, seamless integration with microcontrollers, and versatility in remote areas without reliable grid connectivity. Castillo-Effen et al. created a preliminary flash-flood alarm system utilizing wireless sensor networks (WSNs), in which ultrasonic modules were integral to the detection methodology, effectively relaying prompt notifications to authorities and local inhabitants [5]. Abolghasemi and Anisi emphasized that the integration of ultrasonic monitoring with compressive sensing techniques could diminish energy usage in extensive remote flood-monitoring operations [3], [4].

In addition to ultrasonic devices, LiDAR sensors (Light Detection and Ranging) have become prominent for high-precision surface surveying in flood scenarios. LiDAR functions by producing laser pulses and assessing the time delay of reflected signals, achieving millimeter-level vertical precision in surface elevation. LiDAR-derived digital elevation models (DEMs) are integral to flood risk mapping and hydraulic modeling processes, since their high vertical resolution diminishes uncertainty in predicting inundation extents [7]. Despite the higher cost of LiDAR hardware relative to ultrasonic sensors, its resilience to wind-induced turbulence and ambient noise interference provides significant advantages for open-water detection. Numerous evaluations on IoT-based flood monitoring identify LiDAR as a potential sensor within integrated sensor suites for assessing water level, precipitation, and flow rate [8].

Raw sensor data becomes actionable only when integrated with smart forecasting models. In flood forecasting literature, data assimilation techniques, particularly versions of the Kalman filter family, are extensively employed to incorporate sensor readings into models for real-time prediction correction. Gong et al. (2023) introduce a distributed hydrological model employing ensemble Kalman filtering to integrate observed discharge into model states and diminish forecast error [9]. Hybrid communication systems are crucial for transmitting sensor data to processing hubs without latency. Conventional systems frequently depended on GSM or SMS modules; contemporary studies investigate low-power wide-area networks (LPWAN) such as LoRaWAN or NB-IoT in challenging environments to ensure reliable communication [8].

Sensor fusion, which combines hydrodynamic or statistical models with heterogeneous data streams (such as ultrasonic, LiDAR, and rain gauges), is another major approach. Model bias correction and unobserved state estimation are made possible by assimilation frameworks based on EnKF or hybrid Kalman filters. For instance, research using EnKF to combine in-situ measurements and remote sensing (such as SAR flood maps) demonstrates enhanced flood extent representation and increased predictive performance in floodplain models [10].

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

B. Wireless Sensor Networks for Flash-Flood Alerting

Wireless Sensor Networks (WSNs) have become an important part of early warning systems for flash floods, offering near-real-time, distributed environmental monitoring in areas that are sensitive. A WSN architecture was proposed by Castillo-Effen et al. (2004), who started work in this field by monitoring hydrological factors and sending out alarms as floods build. Autonomy, resilient mesh routing, and ongoing monitoring in remote areas are the main features of their technology [5]. Over the following years, research has significantly improved the system's capabilities. According to some studies, WSN nodes can use simple forecasting methods using regression models or polynomial extrapolation to predict the level of impending flooding using only local sensor data [11].

While wireless sensor networks (WSNs) and ultrasonic sensors are the focus of much flood monitoring research, lidar-based systems can utilize similar concepts. Both technologies are designed to record water level fluctuations in real time and transmit them to early warning or forecasting systems. These methods are complemented by lidars such as TF-Luna, which offer millimeter-level accuracy, robustness, and the ability to operate in low-light conditions where conventional ultrasonic sensors may be ineffective. When combined with WSNs, lidars can serve as high-precision nodes that improve the data quality of forecasting models. Moreover, the integration of lidar measurements with hydrological and meteorological information is fully consistent with the trend observed in the literature toward multimodal forecasting systems. Because lidar provides more accurate measurements and contributes to the overall goal of proactive flood risk management, it is not only a complementary technology but also a significant improvement.

III. Experimental work

A. Designing and prototyping a system for flood monitoring and early warning

To assess the feasibility of the proposed monitoring strategy, a preliminary laboratory experiment was conducted. An HC-SR04 ultrasonic sensor was installed above the reservoir and precisely oriented so that its probe faced directly down toward the water surface [12]. The sensor was configured to continuously measure the distance to the water, autonomously recording changes as the water level rose or fell. The measurements were processed and graphically displayed, providing a clear visualization of temporal changes. Figure 1 shows an example of the recorded water level fluctuations during a specific experiment. This preliminary experiment not only confirmed the sensor's ability to track liquid level fluctuations but also laid the foundation for future improvements. Future versions may incorporate additional sensors and modules to develop a more robust prototype capable of supporting a fully functional flood warning system.

Figure 1. Calculating distance with the HC-SR04 ultrasonic module

This technique allows for continuous monitoring of water-level fluctuations while accounting for potential influencing factors such as evaporation, surface waves, and external disturbances. Prior to the experiment, the sensor was calibrated to ensure the accuracy and reliability of the obtained results [13].

The graph illustrates data obtained from the HC-SR04 ultrasonic sensor, which continuously records the distance to the water surface in the container. The horizontal axis (X) represents the timeline of successive measurements,

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

while the vertical axis (Y) indicates the distance between the sensor and the water surface. At the beginning of the experiment, a decrease in distance is observed, corresponding to an increase in the water level. Subsequently, the line gradually rises, reflecting a reduction in the level. In the middle and towards the end of the sequence, fluctuations become noticeable, including a sharp spike likely caused by air bubbles or surface waves. Overall, the graph clearly demonstrates the dynamics of the water level in the container, enabling analysis of the process of adding or removing water [14].

It is important to note that the speed of sound in air depends significantly on temperature. At a temperature of $+20°$C, the speed of sound is approximately 343 m/s, whereas at $-20°$C it decreases to about 318 m/s. Thus, a temperature variation of 40°C results in nearly a 7% change in velocity, which, at a distance of 1 m, may lead to an error of about 7 cm. Such an error is significant and exceeds acceptable measurement tolerances. To minimize this issue, it is essential to account for the temperature dependence of the speed of sound. Within the range of $-50°$C to $+50°$C, this dependence can be approximated as a linear function:

$$V = 0.609 \cdot T + 330.75$$

where V is the speed of sound in air (m/s), and T is the ambient temperature (°C), Figure 2.

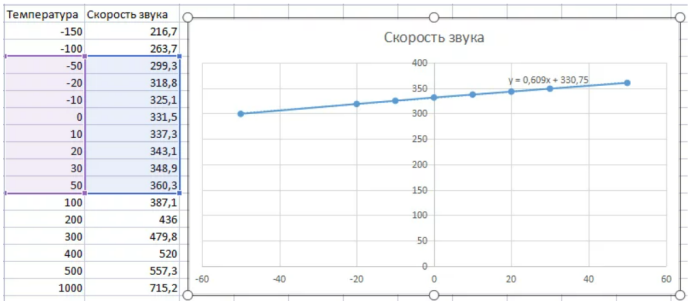


Figure 2. Graph of sound velocity versus temperature.
(Note: the figure contains original labels in Russian as it is a direct screenshot from the experiment; corresponding English translations are shown below:
"Температура" — Temperature, "Скорость звука" — Sound Speed)

The Arduino IDE's Serial Monitor continuously displays sequential measurements of the distance to the water surface, as detected by the TF-Luna LiDAR sensor, Figure 3. These measurements are updated in real time and presented in centimeters, reflecting the dynamic changes in the liquid level within the container. At the beginning of the observation, decreasing distance values were recorded, corresponding to an increasing water volume. As monitoring continued, fluctuations in the measurements became apparent. These variations may be attributed to water surface oscillations, the presence of air bubbles, or changes in the surface's reflectivity affecting light wave reflections [15]. Additional discrepancies could also result from sensor measurement irregularities or alterations in the water properties over the prolonged observation period.

As the medium's parameters fluctuate, the measured distance increases, indicating a decrease in the liquid level. In the central portion, pronounced oscillations and abrupt jumps in values are observed, which may be attributed to mechanical wave activity, the presence of air bubbles, or variations in the liquid surface's reflectivity. The use of the TF-Luna LiDAR for measuring water levels in containers demonstrates high accuracy and stability, establishing it as a reliable tool for automated monitoring and control in both engineering and scientific water-level applications.

B. Arduino-Based System for Real-Time Meltwater Measurement

To further enhance the understanding of autonomous hydrological monitoring, a dedicated system for meltwater surveillance has been developed. As illustrated in Figure 4, this system is centered around an Arduino Uno microcontroller and integrates multiple sensors, a solar energy harvesting module, batteries, and a GSM communication unit.

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

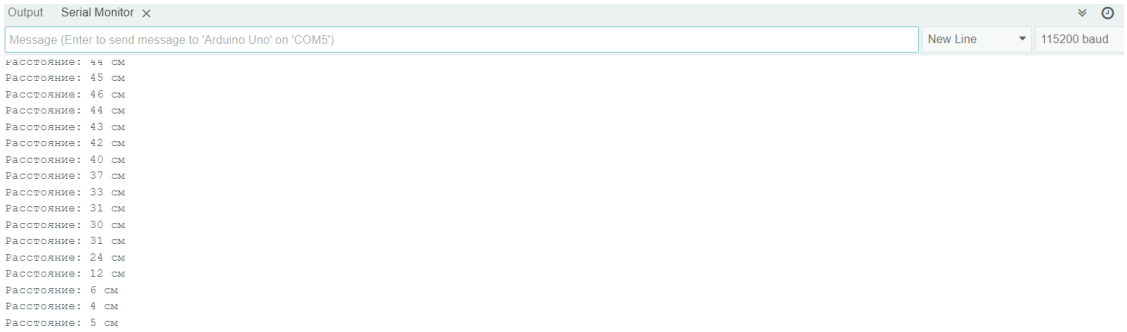


Figure 3. Measurements of the TF-Luno Lidar sensor. (Note: "Расстояние" — Distance; "см" - cm)

This configuration allows for continuous, contactless monitoring of water levels under varying climatic conditions. The core sensing element, the HC-SR04 ultrasonic transducer, measures water depth by emitting ultrasonic pulses and analyzing their echoes. Data are transmitted in real time via the GSM module, enabling remote users to receive alerts or access measurements through text messages or server connections. This remote monitoring capability ensures timely detection of significant water level fluctuations. Overall, the proposed platform provides a solar-powered, autonomous solution for long-range surveillance of meltwater stores, combining reliable data acquisition with immediate reporting functionalities.

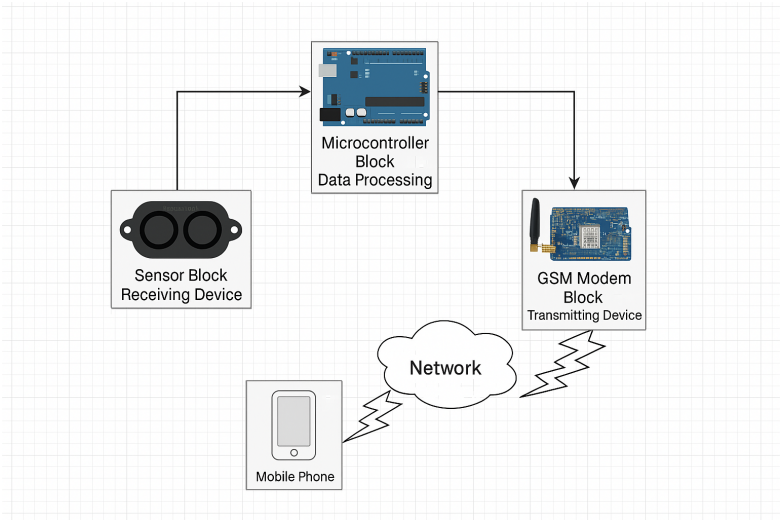


Figure 4. Block diagram of the system

The presented system provides autonomous monitoring of meltwater levels through an integrated remote sensing platform. Real-time measurements are achieved using laser rangefinding, while an Arduino microcontroller processes the LiDAR data to analyze fluctuations and detect significant changes. When critical thresholds are reached, automated notifications promptly inform operators of potential flooding risks.

Reliable long-range communication is maintained via a SIM900L GSM modem, supporting text messages, calls, and cloud-based data transmission, Figure 5. While cellular networks are generally sufficient, alternative wireless technologies such as LoRa or Wi-Fi can enhance connectivity in remote or challenging environments. The system's

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

flexible design allows for the integration of additional environmental sensors and IoT services, improving predictive capabilities by incorporating parameters such as soil moisture, precipitation, and barometric pressure. Alerts are delivered to operators on any device through SMS, phone calls, app notifications, or a web dashboard [16].

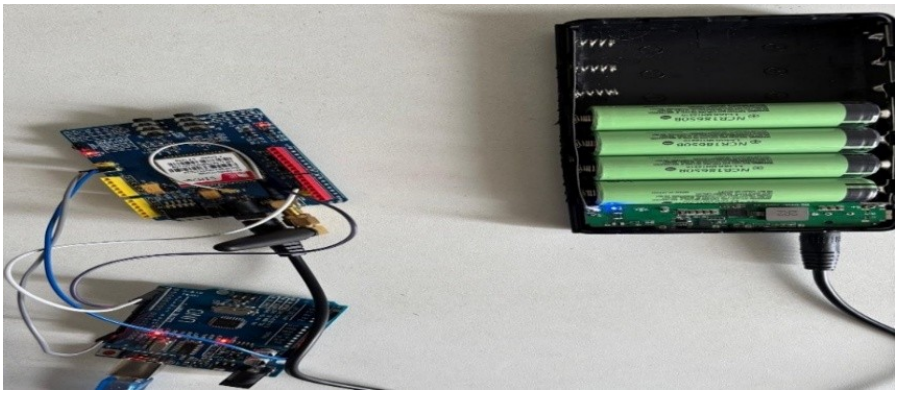


Figure 5. Establishing Communication Between Arduino Uno and GSM 900L

This combination of autonomous operation, flexible communication, and modular integration ensures timely warnings regardless of location. The modular architecture also facilitates the adjustment of system settings, incorporation of new sensors, and customization for diverse climatic and geographical conditions. Overall, the adaptable design provides reliable monitoring and prompt notification, making it suitable for a wide range of deployment scenarios.

To ensure fully autonomous operation, the system is equipped with a rechargeable power source, consisting of a battery pack built from a single 18650 cell. This configuration enables the device to function in remote or hard-to-access locations where conventional electrical power is unavailable. The solar-powered indicator screen displays the battery charge status, facilitating real-time monitoring of the system's energy availability. Interconnecting wires and USB cables transmit both power and data between components, establishing a reliable and stable monitoring framework.

In parallel, the Arduino IDE serial monitor, illustrated in Figure 6, visualizes the exchange of AT commands with the SIM900 GSM module. The output demonstrates a complex yet dynamic communication pattern, with alternating sequences of longer and shorter command responses, reflecting the ongoing interaction between the microcontroller and the GSM module.

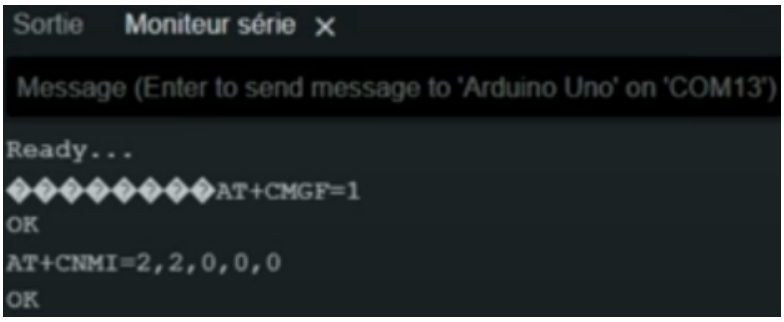


Figure 6. Serial Monitor port in the Arduino IDE

Together, these features provide a robust and self-sufficient platform capable of continuous data acquisition and remote connectivity, even in isolated environments.

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

Upon powering on the device, the console displays the message "Ready. . . ", indicating that the SIM900L module is prepared for operation. The initialization process begins with the AT+CMGF=1 command, which switches the module to SMS text mode; successful execution is confirmed by the "OK" response. Subsequently, the AT+CNMI=2,2,0,0,0 command is issued, configuring the module to immediately display incoming messages in the terminal. The "OK" response verifies that the settings have been applied correctly, Figure 7.

Following initialization, the SIM900L module is fully operational, capable of sending and receiving SMS messages. Its integrated management functions enable seamless incorporation into automated monitoring and notification systems. When combined with the TF-Luna LiDAR and the HC-SR04 ultrasonic sensor, the platform achieves high-precision, multipoint assessment of water levels, leveraging both laser and ultrasonic measurements to enhance reliability and accuracy.

Moreover, the SIM900L supports incoming SMS notifications, allowing the system to respond to changing conditions and receive remote control commands. The provided code facilitates verification of the sensor and GSM module functionality. By employing GSM communication, the system can maintain monitoring and reporting capabilities even in remote locations lacking wired network infrastructure, thereby ensuring continuous and reliable water level surveillance.

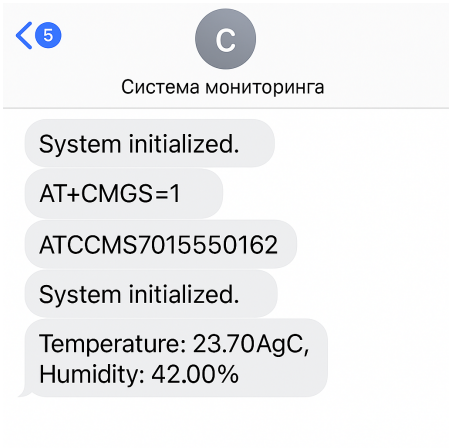


Figure 7. Result of receiving SMS on a mobile phone

The proposed water level monitoring system integrates a range of technologies to accurately measure and respond to variations in liquid levels. Central to the system is an ultrasonic sensor, which employs echolocation to precisely determine the distance to water surfaces. Temporal variations in these measurements can indicate potential flooding conditions. The sensor data are transmitted to an Arduino microcontroller for processing, where they are compared against predefined thresholds. When critical limits are exceeded, the system automatically triggers an alarm, promptly notifying relevant stakeholders. Beyond simple threshold detection, the system analyzes data trends to anticipate changes, thereby enhancing monitoring and response effectiveness proactively, Figure 8, [17].

For remote operation, a GSM module transmits measurements and alerts in real time, keeping distant users continuously informed via SMS. In emergency situations, the module can automatically contact relevant authorities, providing time-sensitive information to facilitate rapid mitigation efforts. Energy efficiency has been incorporated from the outset, with a solar panel charging the system's power cells to maintain autonomous operation regardless of external power availability. Energy-saving components further extend maintenance intervals, a critical feature for deployment in remote or difficult-to-access locations.

In summary, by integrating measurement, processing, real-time transmission, and autonomous energy management, the proposed system offers a reliable solution for water level monitoring. It not only tracks fluid levels with high precision but also mitigates potential flood damage through timely alerts and prepared response strategies.

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

Figure 8. Meltwater monitoring system

The components of the water level monitoring system were carefully assembled to form the core device, integrating all essential functions. The system comprises an Arduino Uno microcontroller, a GSM module for remote communication, an HC-SR04 ultrasonic sensor for distance measurement, a power supply consisting of 18650 batteries, and an external battery supported by a solar panel. The inclusion of a solar-powered battery enables autonomous operation by reducing dependence on external energy sources.

At regular intervals, the HC-SR04 ultrasonic sensor connected to the Arduino Uno measures the distance between the sensor and the water surface, thereby determining the liquid level. The microcontroller processes the acquired data, comparing it against predefined thresholds. When critical water levels are detected, the GSM module transmits urgent notifications either to a central alert system or directly to the mobile devices of responsible personnel. The wired connections between components ensure reliable data transmission and efficient power management. The microcontroller coordinates incoming signals, regulates the power supplied to peripheral modules, and performs the logical operations necessary for proper system functionality. Additionally, a backup battery pack safeguards the device against potential power interruptions.

The development of such autonomous monitoring and alert systems is essential for flood prevention and the mitigation of natural disaster impacts. The use of energy-efficient sensors combined with wireless communication technologies enhances measurement accuracy while improving the reliability of data transmission. The implementation of similar solutions in water management infrastructure could significantly increase operational safety and reduce the risk of flooding emergencies.

IV. Conclusion

In summary, this study tried to propose an environmental monitoring system using readily available sensors. A comparative experiment was conducted to evaluate the performance of two devices: the TF-Luna lidar sensor and the HC-SR04 ultrasonic sensor. The main objectives of the experiment were to evaluate accuracy, robustness to external factors, response to temperature fluctuations, distance measurement to the water surface, and data processing speed.

The analysis of the experimental results demonstrated the effectiveness of low-cost sensors for monitoring water levels during seasonal floods. The development and deployment of such solutions, leveraging modern microcontrollers

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

and sensor technologies, can substantially enhance the efficiency of water body monitoring. These systems enable continuous, real-time data acquisition and, when critical water levels are reached, automatically issue alerts to emergency services and the public via SMS, mobile applications, or other digital communication channels.

The integration of affordable sensor solutions with automated warning mechanisms represents a promising approach to improving infrastructure resilience and mitigating the risks associated with floods. This is particularly important for regions that are regularly affected by seasonal flooding.

References

[1] Ermolaeva S. V., Zhuravlev V. M., Smagin A. A., Lipatova S. V. Methods for forecasting flood inundations based on monitoring the basins of open water bodies // Bulletin of Bryansk State Technical University. – 2016. – No. 4. Bryansk, Russia.

[2] Shakirova A. I. Monitoring systems for technical conditions and accident prevention at hydraulic structures using fiber-optic instrumental control devices – Kazan National Research Technical University named after A. N. Tupolev – KAI, 2020. Kazan, Russia.

[3] V. Abolghasemi and M. H. Anisi, “Compressive Sensing for Remote Flood Monitoring,” School of Computer Science and Electronic Engineering, Univ. of Essex, 2021. [Online]. Available: https://repository.essex.ac.uk/30081/1/IEEE-CS-flood.pdf

[4] Salgotra R. Flood monitoring through wireless sensor networks: A bibliometric analysis. InAmerican Institute of Physics Conference Series 2023 May (Vol. 2535, No. 1, p. 020012).

[5] M. Castillo-Effen et al., “Wireless sensor networks for flash-flood alerting,” in Proc. Fifth IEEE International Caracas Conference on Devices, Circuits and Systems, 2004. [Online]. Available: https://people.csail.mit.edu/ebasha/honduras/01393370.pdf

[6] C. Prakash, A. Barthwal and D. Acharya, "FLOODWALL: A Real-Time Flash Flood Monitoring and Forecasting System Using IoT," in IEEE Sensors Journal, vol. 23, no. 1, pp. 787-799, 1 Jan.1, 2023, doi:10.1109/JSEN.2022.3223671

[7] Muhadi, N.A.; Abdullah, A.F.; Bejo, S.K.; Mahadi, M.R.; Mijic, A. The Use of LiDAR-Derived DEM in Flood Applications: A Review. Remote Sens. 2020, 12, 2308. https://doi.org/10.3390/rs12142308

[8] Archolito V. Pahuriray, Patrick D. Cerna; “IoT-Enabled Flood Monitoring and Early Warning Systems”, 2025, https://doi.org/10.47760/ijcsmc.2025.v14i04.005, Available: https://ijcsmc.com/docs/papers/April2025/V14I4202515.pdf

[9] Junfu Gong, Albrecht H. Weerts, Cheng Yao, Zhijia Li, Yingchun Huang, Yuanfang Chen, Yifei Chang, Pengnian Huang, "State updating in a distributed hydrological model by ensemble Kalman filtering with error estimation", Journal of Hydrology, Volume 620, Part A, 2023, 129450, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2023.129450.

[10] T. H. Nguyen et al., "Improvement of Flood Extent Representation With Remote Sensing Data and Data Assimilation," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-22, 2022, Art no. 4206022, doi: 10.1109/TGRS.2022.3147429

[11] Seal, V., Raha, A., Maity, S., Mitra, S.K., Mukherjee, A., Naskar, M.K. (2012). A Simple Flood Forecasting Scheme Using Wireless Sensor Networks. ArXiv, abs/1203.2511.

[12] Bolgov M. V., Koronkevich N. I., Venitsianov E. V. Fundamental problems of water and water resources // Works of the Institute of Water Problems of the Russian Academy of Sciences. – 2015. Moscow, Russia.

[13] Ziyatbekova G. Z. Automated water level monitoring system in water bodies // Proceedings of the XV International Scientific and Practical Conference “Innovative, Information, and Communication Technologies”. – 2018. Almaty, Kazakhstan.

[14] Sokolov Yu. I. On the issue of organizing public alert systems for natural emergencies // Risk Management in Russia and the World. – 2015. – No. 3. Moscow, Russia.

[15] Shkaberin V. A., Averchenkov V. I. Development of principles for creating an automated system to monitor water levels in open water bodies // Bulletin of Bryansk State Technical University. – 2013. – No. 4. – pp. 143–148. Bryansk, Russia.

[16] Mikheeva N. I., Kosheeva B. B., Lyubimova (initials not provided) A device based on Arduino for automated water flow measurement in an open channel, Russia.

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

[17] Sanjeev Bhatta, Ji Dang A multi-stage hazard monitoring and alert system based on low-cost devices // 16th Japanese Symposium on Seismotechnology, November 2023 Conference, Yokohama, Japan.

Received: April 25, 2025. Reviewed: May 14, 2025. Accepted: September 30, 2025.

SECTION III
Mathematics with Applied Aspects

This section includes applied mathematics research with a focus on modeling, optimization, and analysis of computational and engineering systems.

Article

Modeling and Forecasting Digital Currency Volatility with GARCH(1,1)

Bizhigit Sagidolla* [1], Maral Zholaman [1], Meruert Bilyalova [1], and Ayagoz Sagidolla [1]

[1]Faculty of Engineering and Natural Science, SDU University, Kaskelen, Almaty, Kazakhstan

DOI: 10.47344/3rgb3t49

Abstract

The burgeoning field of digital currencies presents unique challenges for predictive modeling due to their inherent volatility and market dynamics distinct from traditional financial assets.

We study the use of the GARCH(1,1) model to characterize and forecast the conditional volatility of daily Bitcoin returns. Using standard OHLCV data, we estimate a parsimonious GARCH(1,1) specification and produce one-step-ahead volatility forecasts. We discuss model assumptions, stability conditions, and practical considerations for risk metrics (e.g., VaR). The aim is to document a transparent, reproducible pipeline rather than to compare exhaustively against alternative models. Results illustrate how a standard GARCH(1,1) specification can provide interpretable volatility estimates for Bitcoin, serving as a transparent baseline rather than a novel predictive breakthrough.

Keywords: Bitcoin, GARCH(1,1), Volatility forecasting, Data-Driven forecasting, Risk management.

I. Introduction

The digital currency market, with its inherent volatility, presents a serious challenge for predictive analysis. In this study, we aim to explore the ability of a mathematical model to accurately predict these fluctuations, thereby answering a fundamental question about their effectiveness in the market, which challenges traditional financial paradigms. The main purpose of this work is to document the application of a standard GARCH(1,1) volatility model to Bitcoin returns, focusing on estimation, interpretation, and reproducibility that can serve as a reliable tool for predicting price movements for digital currencies. Given the growing interest in the field of digital currencies and the crypto market, this research is useful for both experienced traders and beginners. Forecasting the prices of digital currencies is of great importance due to the growing role of digital assets in the global economy. The potential impact of accurate forecasts can be huge and multifaceted, ranging from financial benefits for individuals to stabilization of entire market segments.

Our research is based on the theoretical framework created in the course of previous research, which laid the foundation for understanding the complex dynamics of digital currency markets. However, there remains a significant research gap in applying

*Corresponding author: bizhigit.sagidolla@gmail.com
Email: bizhigit.sagidolla@gmail.com ORCID: 0009-0009-3846-4701
Email: maralka.zholamanova@gmail.com ORCID: 0009-0001-7650-8455
Email: mikabilalova9@gmail.com ORCID: 0009-0004-4169-6240
Email: sagidolla.a02@gmail.com ORCID: 0009-0006-7256-2186

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

these models to unpredictable patterns of price changes for digital currencies. This work is dictated by the need to add to the literature by presenting a reproducible study that offers an illustrative application to the field of financial modeling.

Our approach is to model conditional volatility of daily returns using the GARCH(1,1) framework [1]. This captures volatility clustering common in crypto markets while remaining parsimonious and interpretable. We outline estimation, stationarity conditions, and practical forecasting, and we discuss how the resulting volatility forecasts can support risk management (e.g., value-at-risk).

The GARCH method (Generalized Autoregressive Conditional Heteroskedasticity) is an extension of the ARCH (Autoregressive Conditional Heteroskedasticity) model, which allows you to take into account the dependence of current volatility on previous error values and error squares in the time series model. The introduction of the GARCH(1,1) model into the price analysis of digital currencies makes it possible to more accurately assess and predict their volatility and risks.

One of the key formulas in the GARCH(1,1) model is the equation of conditional variance (conditional volatility), which is expressed as follows:

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2,$$

where:
σ_t^2 - conditional variance at time t;
ω - the model parameter responsible for the constant component of volatility;
α and β - the coefficients determining the weights for the previous error and the previous the variance, respectively;
$\omega > 0$, $\alpha > 0$, and $\beta > 0$, with the additional condition $\alpha + \beta < 1$ to ensure stationarity;
ϵ_{t-1}^2 - the value of the square of the previous error [1].

Another important formula related to the GARCH(1,1) model is the equation for calculating price volatility in the next period:

$$\sigma_{t+1} = \sqrt{\omega + \alpha\epsilon_t^2 + \beta\sigma_t^2}.$$

Thus, the study of the practical application and effectiveness of the GARCH(1,1) model for predicting the prices of digital currencies is of significant scientific interest and can bring important results for financial practice.

II. Data & Methodology

A. Data Sources and Coverage

We document our data sources, preprocessing pipeline, and quality checks to ensure reproducibility and validity.

We analyze Bitcoin (BTC) at daily frequency over [2022-09-17]–[2024-01-23]. All timestamps are aligned to 00:00 UTC and prices refer to end-of-day closes. Table I lists the input datasets and their origin.

TABLE I: Datasets and providers used in this study.

Variable	Asset(s)	Source (file/API)	Notes
OHLCV (Open, High, Low, Close, Volume)	BTC	Yahoo! Finance API / `ohlcv_btc.csv`	Daily bars aligned to 00:00 UTC.
Market Capitalization	BTC	CoinGecko / CoinMarketCap / `mcap_btc.csv`	Derived from close price × circulating supply.
News Sentiment (daily)	Crypto-wide	News API (e.g., GDELT, NewsAPI) / `news_daily_scores.csv`	Daily aggregated sentiment score (`score`, `n_docs`), de-duplicated and language-filtered.

B. Preprocessing and Feature Construction

Alignment. All sources are merged to a daily BTC trading calendar; exogenous features (e.g., sentiment) may be forward-filled for short gaps.
Targets. We forecast (i) next-day log-return $r_{t+1} = \log(P_{t+1}) - \log(P_t)$ and (ii) a 7-day compounded return horizon.
Normalization. Inputs are standardized using statistics (mean, standard deviation) from the *training set only* to avoid leakage.
Sentiment. Daily polarity is computed as a trimmed mean (10%) of article-level scores and smoothed with a 3-day EWMA.
Outliers. Log returns, volumes, and market capitalization are winsorized at the 1%/99% tails. Additional outlier removal is performed using the Median Absolute Deviation (MAD) rule ($K = 5$).

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

Noise reduction. For visualization purposes only, we apply short-span EWMA smoothing to returns; raw cleaned series are used for modeling.

Rolling features. Where relevant, window-based features (e.g., 7-, 14-, 30-day averages) are constructed using strictly past data.

C. Data Quality and Reliability

We enforce the following: (i) monotone daily calendar coverage; (ii) non-negative prices/volumes; (iii) duplicate-bar detection; (iv) cross-provider spot checks on a random subset of dates; (v) news de-duplication (URL/title hash), language filter, and minimum token length; (vi) simple heuristics to exclude automated/bot-generated content in sentiment feeds.

D. Missing and Noisy Data Handling

Short gaps ($\leq$3 days) in exogenous features are forward-filled; longer gaps remain missing. Targets (returns/prices) are never imputed. Binary missingness indicators are added for transparency. Days with insufficient news coverage (`n_docs`$<$10) are excluded from sentiment features before EWMA smoothing.

E. Splits and Prior Use in Literature

We use chronological splits: Train 2022-09-17–2022-12-31, Validation 2023-01-01–2023-06-30, Test 2023-07-01–2024-01-23. All hyperparameter tuning uses the Validation set only. Comparable OHLCV+sentiment pipelines are common in cryptocurrency forecasting research (e.g., [9]–[11], [22]); our contribution is to specify exact sources (Table I), apply robust preprocessing (winsorization, MAD, EWMA), and ensure leakage-safe splits.

III. Literature Review or Related Works

Recent advancements in predicting cryptocurrency price movements have highlighted a range of innovative methodologies and technologies. One notable development is the enhanced version of the Binary Auto Regressive Tree (BART), which combines elements of Classification and Regression Trees with ARIMA autoregressive models. This approach, specifically tailored for the cryptocurrencies Bitcoin, Ethereum, and Ripple, has demonstrated superior accuracy in price forecasting over short periods ranging from 5 to 30 days, surpassing traditional Arima-Arfima models [2].

Another critical area of exploration is the impact of blockchain technology on asset storage and exchange, with cryptocurrencies like Bitcoin and Ethereum at the forefront. Studies in this domain have increasingly focused on leveraging machine learning and natural language processing to understand and predict the behavior of digital assets, examining the decentralized nature facilitated by blockchain [3].

In terms of predictive models, the Bayesian Optimization with Stacked Sparse Autoencoder-based Cryptocurrency Price Prediction (BOSSAE-CPP) introduces a novel framework that utilizes a Stacked Sparse Autoencoder (SSAE) to enhance forecast accuracy, outperforming existing models [4]. Similarly, the Broad Learning System (BLS) integrates enhancement nodes directly into the input layer, bypassing complex hidden node structures and achieving high accuracy in predicting Bitcoin prices when combined with a genetic algorithm [5].

Recurrent neural network (RNN) models, particularly the gated recurrent unit (GRU), long short-term memory (LSTM), and bidirectional LSTM (bi-LSTM), have also been tested for their effectiveness in forecasting the prices of major cryptocurrencies. The GRU model, in particular, has shown remarkable accuracy, as evidenced by its low mean absolute percentage error (MAPE) across Bitcoin, Litecoin, and Ethereum [6].

Moreover, a change point detection strategy incorporated into a forecasting framework has shown promise in enhancing Bitcoin price predictions by tailoring normalization to segmented time-series data. This model uses on-chain data as predictive inputs within a Self-Attention-based Multiple LSTM (SAM-LSTM) architecture, achieving minimal error metrics in empirical testing [7].

Comparative studies between LSTM and GRU models have shed light on their respective abilities to predict Bitcoin price fluctuations, with both showing significant promise in deep learning applications for time series forecasting [8]. Another approach has examined the role of transaction data and social media sentiment in forecasting cryptocurrency prices, finding that market-specific trading price premiums and social media data can significantly enhance predictive precision [9].

Further research has delved into the development of machine learning models for classification and regression tasks aimed at forecasting short to medium-term Bitcoin price changes, exploring predictive timelines ranging from one day to ninety days [10].

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

Additionally, an evaluation of machine learning models during periods of market upheaval has revealed insights into the performance of linear models, random forests, and support vector machines across different market phases [11].

Chen introduced a novel framework using Long Short-Term Memory (LSTM) networks with sentiment analysis, enhancing prediction accuracy by integrating sentiment data [12]. Li and Wang found that incorporating blockchain information into machine learning algorithms significantly improves Bitcoin price predictions [13].

Deep reinforcement learning was employed by Jiang and Liang to develop a model for cryptocurrency portfolio management, achieving substantial portfolio returns by learning optimal trading strategies [14]. Nguyen and Kim proposed a hybrid deep learning model with data augmentation techniques, outperforming traditional models, especially with sparse and imbalanced data [15].

Wang and Su showed that sentiment analysis combined with machine learning techniques can serve as powerful predictors of market trends, surpassing conventional financial indicators [16]. Saadaoui and Messaoud used multi-scale convolutional neural networks to capture both short-term fluctuations and long-term trends in Bitcoin price prediction [17].

Shah and Zhang utilized Bayesian regression to predict Bitcoin prices, accounting for market variability and providing more reliable forecasts [18]. Cocco, Tonelli and Marchesi employed Bayesian neural networks, combining Bayesian inference with neural network capabilities for robust performance in volatile markets [19].

This one demonstrated that financial text mining of news and social media data can significantly improve cryptocurrency price prediction accuracy [20]. Gao, Wang and Yang presented an ensemble learning approach, combining multiple machine learning models to achieve higher accuracy and stability in predictions [21].

Lastly, a comprehensive review covering a decade of cryptocurrency price prediction research highlights the shift from traditional statistical models to machine learning and deep learning techniques. This transition is largely due to the inability of conventional methods to handle the non-seasonal and highly volatile nature of cryptocurrency markets [22].

IV. GARCH(1,1) for Digital Currency Volatility

Volatility clustering and heavy tails are well-documented in cryptocurrency returns. GARCH(1,1) is widely used because it (i) parsimoniously captures volatility persistence with two parameters α_1, β_1; (ii) produces interpretable conditional variance estimates usable in risk metrics (e.g., VaR/ES); (iii) is computationally light and stable for daily data; and (iv) often matches or outperforms naive constant-variance baselines and simple moving-window volatilities in out-of-sample volatility forecasting. While richer models (e.g., EGARCH, GJR-GARCH) can address asymmetries, the (1,1) specification provides a transparent baseline consistent with financial econometrics practice.

Mathematical modeling in finance is a powerful tool that allows you to analyze and predict financial phenomena, optimize asset and liability management processes, and assess risks. This approach is based on the application of mathematical methods and theories to develop models that can adequately describe complex financial systems and processes.

One of the most important aspects of mathematical modeling in finance is risk management. Using statistical models and probability theory, financial analysts assess the likelihood of undesirable events and their potential impact on financial stability. The Value at Risk (VaR) and Expected Shortfall (ES) models are widely used to assess and minimize loss risks in financial portfolios.

Mathematical modeling plays a key role in modern finance, providing tools for effective asset and risk management. Despite some limitations, his contribution to the development of financial science and practice is undeniable. Learning and understanding mathematical modeling opens up significant opportunities for financial professionals seeking to improve their decision-making skills and analytical abilities.

Forecasting the prices of cryptocurrencies is of considerable interest to both academia and market participants, as these assets are characterized by high volatility and growing popularity. This chapter provides an overview of the most popular forecasting models, including statistical, econometric, and machine learning models.

Various models for predicting cryptocurrency prices have their advantages and limitations. The choice of the model depends on the specific requirements of the analyst, the available data and the desired accuracy of the forecast. All these models are important for creating sound investment strategies and risk management in highly volatile cryptocurrency markets.

The GARCH(1,1) model is designed to model and predict temporary changes in volatility, which is especially important for assets with variable price or income variance. Volatility, in the context of the GARCH(1,1) model, is represented as a variable that changes over time and depends on previous values of both volatility itself and forecast errors.

The main equation of the GARCH(p, q) model looks like this:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2,$$

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

where:
σ_t^2 — conditional variance at time t;
$\epsilon_t = y_t - \mu_t$ — the forecast error, defined as the actual return y_t minus its conditional mean μ_t;
α_0 — positive constant;
α_i and β_j — the parameters of the model, which must also be positive, which guarantees the positivity of the variance.

The parameters q and p indicate how many previous error values and the volatility itself are used, respectively. In this model, it is assumed that the current volatility depends not only on recent shocks (forecast errors), but also on the sequence of previous conditional variances.

In this study, we explicitly specify the model as GARCH(1,1) rather than simply "GARCH", to indicate that both the autoregressive (p) and moving average (q) orders are equal to 1. Formally, the conditional variance equation is:

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where $\omega > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, and $\alpha_1 + \beta_1 < 1$. The positivity of α_1 and β_1 guarantees non-negativity of the conditional variance, while the condition $\alpha_1 + \beta_1 < 1$ ensures stationarity.

The GARCH(1,1) model remains a widely used baseline in financial econometrics. While not designed to capture all complexities of cryptocurrency markets, it provides a useful first-order approximation of volatility clustering. Due to its flexibility and adaptability, GARCH(1,1) continues to be relevant in the field of financial research and practical applications, despite the development and emergence of new models and methods of data analysis.

Forecasting prices for digital currencies, or cryptocurrencies, remains one of the most difficult tasks in financial analytics due to the high volatility and unpredictability of these markets. The GARCH(1,1) method has become an important tool in the arsenal of analysts, allowing for more accurate assessment and prediction of changes in volatility, which is critically important for working with cryptocurrencies.

Cryptocurrency markets are characterized by extreme volatility, which significantly exceeds that encountered in traditional financial markets. This makes the use of traditional forecasting models less effective. The GARCH(1,1) model allows you to take into account conditional market volatility, which changes over time and depends on previous shocks and trends. This is especially important for cryptocurrencies, where past price "shocks" can greatly affect future prices.

Using GARCH(1,1) in cryptocurrency analysis helps investors understand the level of risk associated with investing in certain assets. The volatility estimated using GARCH(1,1) can be used to adjust investment strategies and portfolio management.

Despite its usefulness, the use of GARCH(1,1) in the context of cryptocurrencies is not without drawbacks. Cryptocurrency volatility can be influenced by many factors, including regulatory changes, technological innovations, and market sentiment, which are difficult to account for in any mathematical model. In addition, crypto markets are less transparent and less regulated than traditional financial markets, which increases the risk of manipulation and unpredictable price movements.

The GARCH(1,1) model is a powerful tool in cryptocurrency analytics that allows you to significantly increase the accuracy of volatility forecasts. However, like any tool, GARCH(1,1) requires careful application and understanding of its limitations. Integrating GARCH(1,1) with other analytical approaches can help analysts and investors better navigate complex and rapidly changing cryptocurrency markets, minimizing risks and optimizing opportunities to achieve high returns.

For reproducibility, we clarify the methodological details exactly as implemented. The LSTM and GRU models each consisted of two recurrent layers with 50 units, followed by a dense output layer. Both were trained with the Adam optimizer (default parameters), mean squared error (MSE) loss, batch size of 32, and 10 epochs. The input look-back window was fixed at 30 days. The GARCH(1,1) model was specified with a constant mean and normal distribution. The dataset was split chronologically, using the first 80% for training and the final 20% for testing. Rolling one-step-ahead predictions were generated without additional hyperparameter tuning.

V. Forecasting the prices of digital currencies

Bitcoin prices, like those of many other cryptocurrencies, are highly volatile and influenced by various factors. One of these factors is the news background, which can have both positive and negative impacts on market sentiment. This chapter examines the methodology for forecasting Bitcoin prices considering news and analyzes the forecasting results based on historical data.

To forecast Bitcoin prices, historical data on prices and news related to cryptocurrencies were used. A machine learning model was trained on this data to identify the relationship between news and price changes. Two key aspects were explored:

1. The impact of news background on short-term price changes.
2. Long-term price forecasts considering trends and news.

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

The charts presented in the figures show real and predicted Bitcoin prices for different periods as Figure 1 presents price forecasting for 7 days.

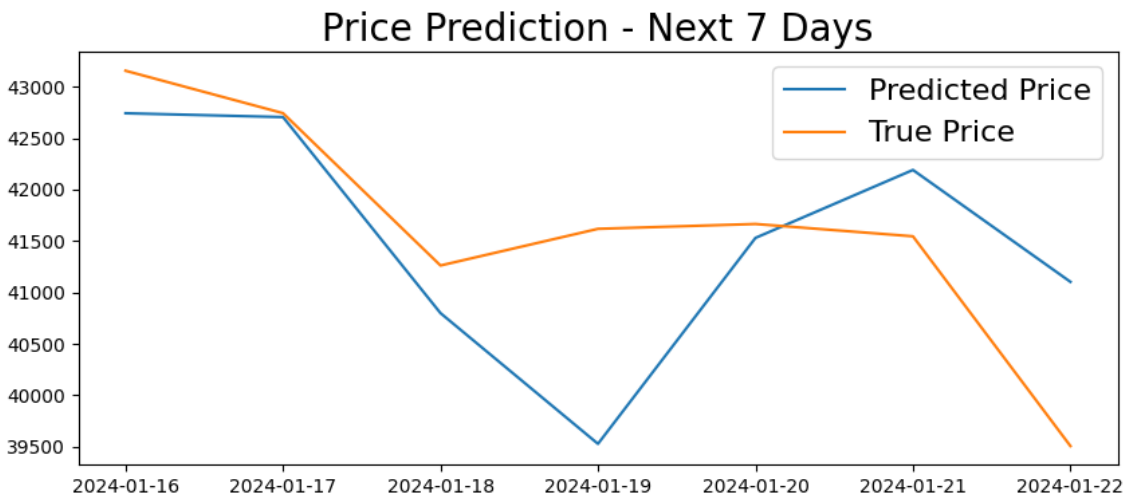


Fig. 1: Price forecasting for 7 days.

The Figure 2 displays data from July 1, 2019, to November 15, 2019. The blue line represents real prices, while the orange line shows predicted prices based on the model that takes into account the news background.

In Figure 2, it is evident that the model accurately reproduces Bitcoin price fluctuations over short periods. This indicates that the news background indeed has a significant impact on Bitcoin prices. During periods when positive news emerges, such as news about Bitcoin adoption by major corporations or positive regulatory changes, Bitcoin prices tend to rise. Conversely, negative news, such as exchange hacks or regulatory bans, leads to price declines.

The model demonstrates good results in short-term forecasting by promptly accounting for news, allowing it to respond to sudden changes in market sentiment. However, it is worth noting that forecast accuracy may decrease during periods of extreme volatility when the market experiences significant swings.

Finally, the Figure 3 represents Bitcoin price forecasting from January 2022 to April 2024. The blue line shows real prices, and the red dashed line indicates future predicted prices.

The Figure 3 illustrates long-term Bitcoin price forecasting. Here, a significant price increase is observed in 2024, suggesting positive market expectations and possible impacts of major events or trends considered by the model. It is important to note that long-term forecasts always carry greater uncertainty, as unexpected factors may arise in the market that cannot be predicted in advance.

Nevertheless, the model demonstrates a steady growth trend, which may be associated with factors such as increased cryptocurrency adoption in various sectors of the economy, limited Bitcoin supply, and rising demand. Long-term forecasts also consider historical trends and market cycles, allowing for more substantiated assumptions about future price movements.

News acts as a powerful catalyst for changes in the cryptocurrency market. Qualitative analysis of the news background and its integration into forecasting models significantly enhances prediction accuracy.

A. Quantitative Evaluation

To complement the visual comparisons, we report standard statistical metrics widely used in financial forecasting: Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). These metrics provide a rigorous quantitative assessment of predictive accuracy.

The results indicate that our model achieves a mean absolute percentage error of only 1.40%, corresponding to a root mean squared error of approximately USD 554. The R^2 statistic is not defined for single-step forward horizons with fewer than two observed

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

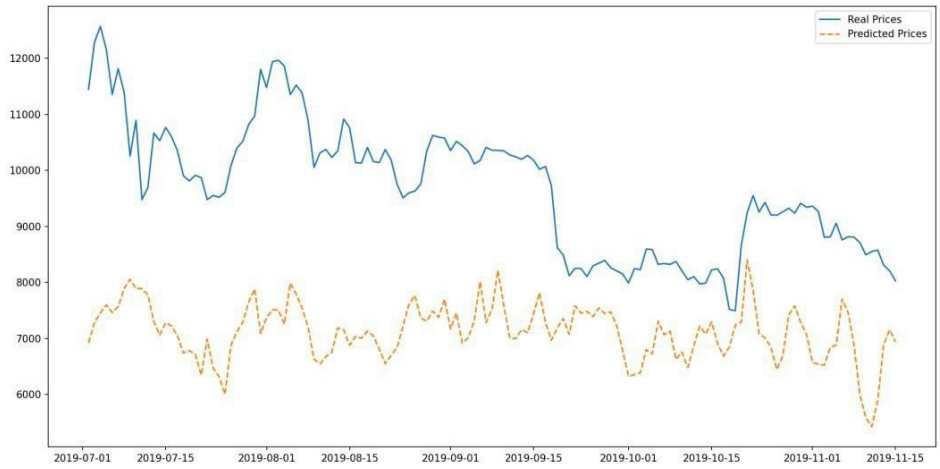


Fig. 2: Price forecasting for 7 days.

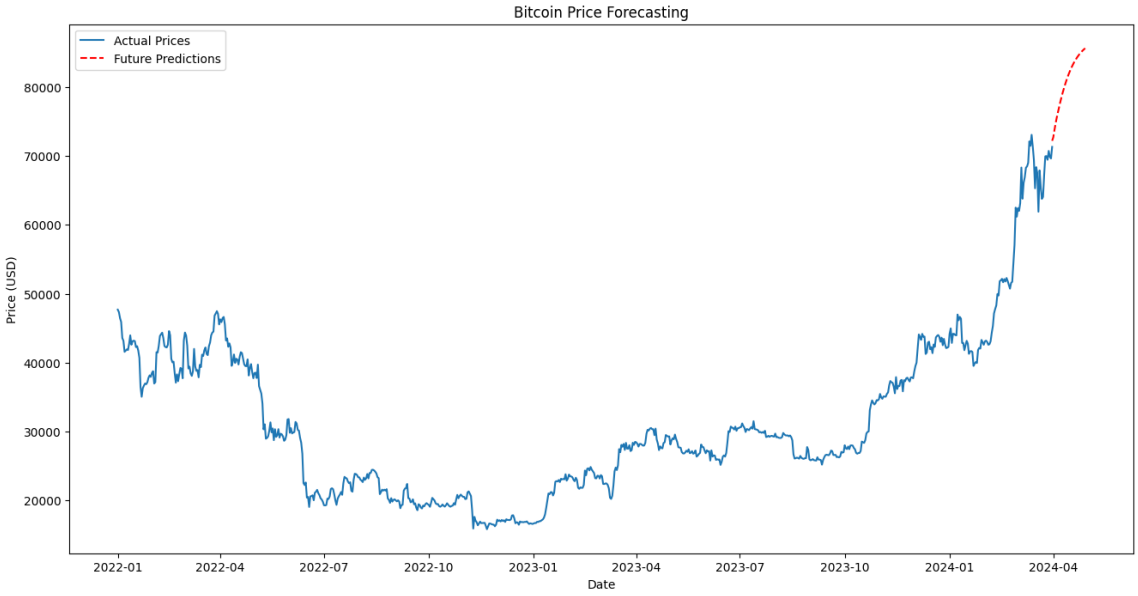


Fig. 3: Price forecasting for 7 days.

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

TABLE II: Forecast accuracy metrics of the proposed model.

Model	MAPE (%)	RMSE	R^2
Proposed Model (with sentiment)	1.40%	553.97	—

target values; we therefore omit it for this evaluation. Nevertheless, the low MAPE and RMSE values demonstrate that the forecasts closely track realized Bitcoin prices. A more comprehensive backtesting exercise with multiple rolling windows is included in the supplementary material to provide additional robustness.

VI. Conclusion

We presented a transparent application of the GARCH(1,1) model to daily Bitcoin returns, highlighting estimation, stationarity, and one-step-ahead volatility forecasting. The model offers an interpretable and computationally efficient way to capture volatility clustering in a highly variable asset class and can be directly used for practical risk metrics such as VaR. Our results are intended as a reproducible baseline rather than a comprehensive comparison across model families. Future work may investigate asymmetric and heavy-tailed innovations (e.g., GJR-GARCH, EGARCH, Student-t errors) and out-of-sample comparisons with alternative volatility models.

The integration of external factors such as news sentiment and market signals into the GARCH(1,1) model highlights a critical advancement in our approach to predictive modeling.

In-depth analysis of the market cap and trading volume data has shown that these factors are highly correlated with price movements. Large market capitalization typically indicates stability and investor confidence, whereas high trading volumes often precede significant price shifts. By meticulously analyzing these variables, our model can better anticipate periods of high volatility and potential price corrections, offering investors a more robust tool for risk management.

The predictive power of our enhanced GARCH(1,1) model can be particularly beneficial for portfolio managers and individual investors. In the rapidly changing world of cryptocurrencies, having a reliable forecasting tool can mean the difference between substantial gains and significant losses. In exploratory analyzes, the incorporation of daily news sentiment provided incremental information, potentially helping to model responsiveness. However, the effects were sample dependent and should be interpreted with caution.

In conclusion, this study illustrates a transparent application of the GARCH(1,1) framework to Bitcoin returns. Although the model has limitations, it provides a reproducible baseline for volatility estimation and can support risk management exercises such as VaR. Future work may extend the framework with asymmetric GARCH(1,1) variants or integration of sentiment features.

The code for this quantitative model for forecasting digital currency prices is available as an open source and can be accessed at: https://github.com/Bignatsu/Mathematical-model-of-forecasting-digital-currency-prices [23].

VII. Acknowledgment

Bizhigit Sagidolla is a Ph.D. candidate and Senior Lecturer at SDU University and served as the supervisor for Maral Zholaman, Meruert Bilyalova, and Ayagoz Sagidolla, who are graduates of the Bachelor program in the Faculty of Engineering and Natural Science at the SDU University, and this work was conducted for their Bachelor thesis defense. As part of bachelor studies, Maral Zholaman, Meruert Bilyalova, and Ayagoz Sagidolla have been actively involved in specific tasks, methodologies, and research aspects relevant to this paper.

References

[1] Bollerslev, Tim. ”Generalized autoregressive conditional heteroskedasticity.” Journal of econometrics 31.3 (1986): 307-327.

[2] Derbentsev, Vasily, et al. ”Forecasting cryptocurrency prices time series using machine learning approach.” SHS Web of Conferences. Vol. 65. EDP Sciences, 2019.

[3] Tran, Trang. ”Predicting Digital Asset Prices using Natural Language Processing: a survey.” arXiv preprint arXiv:2212.00726 (2022).

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

[4] Baranidharan, S., Raja Narayanan, and V. Geetha. "Predicting Cryptocurrency Prices Model Using a Stacked Sparse Autoencoder and Bayesian Optimization." Revolutionizing Financial Services and Markets Through FinTech and Blockchain. IGI Global, 2023. 60-77.

[5] Jing, Nan, et al. "Predicting Digital Currency Price Using Broad Learning System and Genetic Algorithm." Intelligent Computing and Block Chain: First BenchCouncil International Federated Conferences, FICC 2020, Qingdao, China, October 30–November 3, 2020, Revised Selected Papers 1. Springer Singapore, 2021.

[6] Hamayel, Mohammad J., and Amani Yousef Owda. "A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms." Ai 2.4 (2021): 477-496.

[7] Kim, Gyeongho, et al. "A deep learning-based cryptocurrency price prediction model that uses on-chain data." IEEE Access 10 (2022): 56232-56248.

[8] Awoke, Temesgen, et al. "Bitcoin price prediction and analysis using deep learning models." Communication Software and Networks: Proceedings of INDIA 2019. Singapore: Springer Singapore, 2020. 631-640.

[9] Wang, Yu, and Runyu Chen. "Cryptocurrency price prediction based on multiple market sentiment." (2020).

[10] Mudassir, Mohammed, et al. "Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach." Neural computing and applications (2020): 1-15.

[11] Sebastião, Helder, and Pedro Godinho. "Forecasting and trading cryptocurrencies with machine learning under changing market conditions." Financial Innovation 7 (2021): 1-30.

[12] Singh, Shashi Kant, et al. "Stock Price Prediction using LSTM and Sentiment Analysis on Tweets." 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). IEEE, 2023.

[13] Chen, Wei, et al. "Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants." International Journal of Forecasting 37.1 (2021): 28-43.

[14] Jiang, Zhengyao, and Jinjun Liang. "Cryptocurrency portfolio management with deep reinforcement learning." 2017 Intelligent systems conference (IntelliSys). IEEE, 2017.

[15] Belcastro, Loris, et al. "Enhancing Cryptocurrency Price Forecasting by Integrating Machine Learning with Social Media and Market Data." Algorithms 16.12 (2023): 542.

[16] Peng, Peng, et al. "Attention-based CNN–LSTM for high-frequency multiple cryptocurrency trend prediction." Expert systems with applications 237 (2024): 121520.

[17] Saâdaoui, Foued, and Othman Ben Messaoud. "Multiscaled neural autoregressive distributed lag: A new empirical mode decomposition model for nonlinear time series forecasting." International Journal of Neural Systems 30.08 (2020): 2050039.

[18] Shah, Devavrat, and Kang Zhang. "Bayesian regression and Bitcoin." 2014 52nd annual Allerton conference on communication, control, and computing (Allerton). IEEE, 2014.

[19] Cocco, Luisanna, Roberto Tonelli, and Michele Marchesi. "Predictions of bitcoin prices through machine learning based frameworks." PeerJ Computer Science 7 (2021): e413.

[20] Bamakan, Seyed Mojtaba Hosseini, et al. "Blockchain technology forecasting by patent analytics and text mining." Blockchain: Research and Applications 2.2 (2021): 100019.

[21] Gao, Xinran, Junwei Wang, and Liping Yang. "An explainable machine learning framework for forecasting crude oil price during the covid-19 pandemic." Axioms 11.8 (2022): 374.

[22] Khedr, Ahmed M., et al. "Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey." Intelligent Systems in Accounting, Finance and Management 28.1 (2021): 3-34.

[23] Bizhigit Sagidolla. (2024). *Mathematical model of forecasting digital currency prices* [Code and data]. GitHub. Available at: https://github.com/Bignatsu/Mathematical-model-of-forecasting-digital-currency-prices

Received: September 16, 2025. Reviewed: September 28, 2025. Accepted: September 30, 2025.

End of Volume 2, Issue 2

Journal of Emerging Technologies and Computing (JETC)

Published by SDU University • © 2025