BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 14, 2026
Research Article

Data Pre-processing Approach with ML Algorithms for Accuracy and Authenticity Detection of Big Data Sourced from the Social Internet of Things and a Case Study

Deniz Kizilaslan[1], Sukru Mustafa Kaya[2]

***Abstract*— The Social Internet of Things (SIoT), which integrates sensor data with social media interactions, produces massive volumes of unstructured data that require accurate and reliable preprocessing for meaningful analysis. This study investigates the effectiveness of various machine learning (ML) classification algorithms in detecting the accuracy and authenticity of SIoT-derived data.A dataset comprising 17,500 user records collected from mobile devices and social media platforms was analyzed using five ML classifiers: Logistic Regression, Naive Bayes, K-Nearest Neighbor (K-NN), Random Forest, and Support Vector Machines (SVM). Through extensive hyperparameter tuning and 5-fold cross-validation, the Random Forest and SVM models exhibited the highest performance, achieving accuracy scores of 0.58 and 0.57, respectively. SVM also obtained the best AUC value of 0.64, highlighting its strength in distinguishing authentic from manipulated data. Additionally, the results emphasize the need for larger, more diverse datasets, and suggest incorporating deep learning and automated bias mitigation methods in future research.**

Index Terms—**Big Data, Classification Algorithms, Data Processing, Internet of Things, Machine Learning**

I. INTRODUCTION

In the age of digital change, IoT is a very important technologıcal trend today. The number of IoT devices is expected to increase by 25.1 billion by 2025. A relatively new concept in the IoT family, a subset called SIoT (Internet of Social Things), is a method of integrating IoT with social networks. SIoT is a simulation of human-to-human and object-to-object social networks, where humans are referred to as intellectual and relational objects. They build their social networks to achieve common goals, such as increasing accessibility, success, and productivity, as well as providing the services they need [1]. Along with the rapid developments in the Internet of Things, the increasing accessibility of ubiquitous computing resources and mobile devices, the prevalence of rich media contents, and the subsequent social, economic and cultural changes, computer technology and applications have also been rapidly developing and changing in recent years. The proliferation of social relationships between IoT entities goes beyond personal computers. The increase in the number of these relationships and their heterogeneity lead to a number of problems. Different studies are being carried out to solve these problems. This study uniquely contributes by systematically evaluating the accuracy and reliability of IoT-based social media data using machine learning techniques, aiming to enhance trust and reduce misinformation in social media platforms. This novel approach supports increased confidence in digital communication channels and promotes more reliable social ecosystems. This study discusses the role of IoT in social media, examines the management of relationships, the problem of social relationships and the rapid development in IoT and proposes solutions using machine learning and deep learning techniques [2]. The Internet of Social Things (SIoT) seeks to overcome the challenges of the Internet of Things (IoT), such as scalability, trust, and resource discovery, inspired by the processing of social data. This study aims to examine research on SIoT from two perspectives. Application domain and integration into new computing models. For this, a two-dimensional study layer is proposed and projects are investigated accordingly. The first dimension considers and classifies existing research from the perspective of the application field; The second dimension does the same in terms of integration into new computing models. The aim of this study is to define SIoT technically, classify related research, encourage the dissemination of current technologies, and contribute to the literature by discussing the studies conducted in this field [3]. This study focuses on the pioneering impact of the social internet of things (S-IoT), especially in the political economy. The study defines S-IoT, determines its scope, and analyzes the shaping and guiding influences within the scope. It is analyzed how and in what direction the second effect shapes individual preferences in political economy. Rather than a scientific and technological development, the Internet of Things is both a pioneer of the digital revolution and almost the digital revolution itself. S-IoT and its enormous impact on our lives and

[1] **Deniz Kızılaslan**, is with Department of Computer Engineering University of Istanbul Aydin University, Istanbul, Turkey, (e-mail: denizkizilaslan@stu.aydin.edu.tr). https://orcid.org/0009-0009-7128-4075
[2] **Şükrü Mustafa Kaya**, is with Department of Computer Technologies, Blockchain Application Research Centre, University of Istanbul Aydin University, Istanbul, Turkey, (e-mail: mustafakaya@stu.aydin.edu.tr). https://orcid.org/0000-0003-2710-0063
Manuscript received Jun 3, 2025; accepted Dec 3, 2025. DOI: 10.17694/bajece.1712376

Kızılaslan, D., Kaya, S.M. (2026). Data Pre-processing Approach with ML Algorithms for Accuracy and Authenticity Detection of Big Data Sourced from the Social Internet of Things and a Case Study. Balkan Journal of Electrical and Computer Engineering, 14, 63-73.

choices is therefore the main focus of this article. Our conclusion is that the free will of the electorate has come to the point of

extinction through the social internet of things [4]. The primary aim of a different study is to present a comprehensive review article from the SIoT system to analyze and evaluate recent studies in this field. Therefore, the study focuses on the main components (Architecture, Relationship Management, Trust Management, web services and information), features, parameters and challenges of SIoT. Articles published between 2011 and December 2019 are reviewed to collect sufficient information for better analysis. The strengths and weaknesses of each article are examined, and effective evaluation parameters, approaches, and simulation tools most used in this field are discussed. For this purpose, a scientific classification for the final SIoT structure is presented based on the academic studies we have reviewed [5].

In our age where digitalization increases day by day, the concept of Internet of Things (IoT) plays an important role in transforming the digital world by creating an ecosystem where objects can communicate with each other and share data over the network. Big data collected through IoT devices causes various interactions, especially on social media platforms. However, the problems of accuracy and reliability of the information contained in large data sets emerge as one of the most important problems of our age.

This study will focus on Machine Learning (ML) algorithms used to evaluate the accuracy and reliability of big data originating from IoT on social media platforms. By analyzing patterns in large data sets, ML algorithms can provide valuable insight into the authenticity of the information they contain. The aim is to detect false or manipulated information spread on social media and use ML algorithms effectively in this context.

The study will primarily highlight the accuracy and reliability problems in big data originating from IoT. Then, through ML algorithms, it is aimed to evaluate the information found and shared on social media with the help of algorithms and determine their accuracy levels. In addition, the study adds innovation to the real and reliable social platforms by contributing to reliable digitalization and increasing reliability both individually and collectively in terms of filtering the data widespread on social platforms and reflecting the facts.

This study aims to make a valuable contribution to research in this field by providing an effective methodology to increase the accuracy and realism of large data sets generated by IoT. These methods aim to obtain accurate and reliable data and to make a comprehensive evaluation of the performance of classification algorithms.

The main contributions of this study are as follows:

- It presents a systematic approach to assessing the accuracy and authenticity of large-scale data originating from IoT-based social media platforms (SIoT).
- It compares the performance of various machine learning classification algorithms (Naive Bayes, Random Forest, and Support Vector Machine) on real-world data for detecting unreliable or manipulated content.
- It provides an evaluation framework based on performance metrics such as accuracy, precision, recall, F1-score, and AUC to assess model effectiveness.
- It contributes to the development of more reliable digital ecosystems by enhancing data filtering and trust management capabilities in.

II. LITERATURE REVIEW

The Internet of Social Things (SIoT) and the accuracy of the big data obtained from this domain have attracted increasing attention in recent years. SIoT integrates the Internet of Things with social networks, modeling interactions between objects through social structures [6]. In this context, the accuracy and reliability of large datasets collected from SIoT are critically important for data-driven applications [7].

Machine learning algorithms are widely used to assess the accuracy and authenticity of big data sources. In particular, classification methods play an effective role in detecting erroneous, fake, or manipulated information within datasets [8]. In this field, algorithms such as Naive Bayes, Random Forest, and Support Vector Machines (SVM) have been compared in terms of accuracy and performance on various datasets.

Recently, research focusing on the accuracy of big data has increasingly concentrated on social media data derived from IoT sources. Zhang et al. (2021) presented methods aimed at enhancing user experience through the integration of social media and IoT, emphasizing semantic integration and personalization techniques [9]. Meanwhile, Kim et al. (2020) focused more on analyzing temporal interaction patterns and anomaly detection within social platforms [10]. Chen et al. (2019) highlighted the role of trust management and reputation systems in improving the reliability of SIoT environments [11]. Similarly, Li and Wang (2020) proposed a hybrid machine learning framework for sentiment-aware social IoT applications, demonstrating how emotional context can enhance prediction accuracy [12]. Furthermore, Gupta et al. (2022) explored real-time social IoT analytics, showing the potential of edge computing in reducing latency and improving decision-making accuracy [13]. While both studies contribute valuable insights, their differing approaches highlight the multifaceted nature of SIoT data accuracy challenges.

However, comprehensive and comparative studies on the accuracy of big data originating from SIoT remain limited. Therefore, this study provides a significant contribution to the literature by presenting a comparative performance evaluation of multiple machine learning algorithms for detecting accuracy and authenticity in large datasets obtained from SIoT.

Table I. Comparative analysis and literature gap.

Study	Data Source / Focus	Method	Strengths	Limitations	Contribution of This Study
Zhang et al. (2021)	Social media + IoT integration for user experience	Semantic integration, personalization	Enhanced personalization and semantic enrichment	Limited focus on scalability and anomaly detection	This study addresses broader SIoT data accuracy issues beyond personalization.
Kim et al. (2020)	Temporal interaction patterns & anomaly detection	Temporal analysis, anomaly detection	Improved detection of unusual behaviors, increased reliability	Did not incorporate user sentiment or trust perspectives	This integrate trust and sentiment aspects along with anomaly detection.
Chen et al. (2019)	Trust & reputation in SIoT	Trust management frameworks	Increased reliability of SIoT interactions	Lack of experimental validation with large-scale real-world datasets	We validate trust-related insights with empirical data and combine them with predictive models.
Li & Wang (2020)	Sentiment-aware social IoT	Hybrid ML, sentiment analysis	Demonstrated how emotional context improves prediction accuracy	Narrow focus on sentiment only, missing real-time processing	We extend sentiment analysis with real-time and large-scale SIoT data handling.
Gupta et al. (2022)	Real-time social IoT analytics with edge computing	Edge computing, real-time analytics	Reduced latency, improved decision-making	Limited exploration of accuracy and trust dimensions	We combine edge-based analytics with accuracy and trust evaluation in SIoT data.
This Study	Accuracy of SIoT big data (social media sources)	ML-based prediction + comparative analysis	Comprehensive evaluation of multiple ML models with accuracy & performance	Needs further exploration of deep learning models (future work)	Provides holistic accuracy analysis, comparative evaluation, and highlights gaps in literature.

These studies provide valuable insights into SIoT data processing, trust evaluation, and anomaly detection. However, the majority of prior work either focuses on social relations, trustworthiness, or anomaly detection, without systematically comparing multiple machine learning algorithms on SIoT datasets for accuracy and authenticity. In contrast, the present study addresses this gap by providing a comparative analysis of multiple ML algorithms, highlighting their performance in detecting inaccurate or manipulated data in large-scale SIoT datasets.

III. INTERNET OF THINGS AND BIG DATA

Technology in the study has undergone a significant evolution today. The amount of data produced instantly has reached very high levels, which makes it difficult to process, understand and transform data into information. New techniques and methods have been developed to report this high amount of data and transmit it to the end user. In particular, the increasing use of Internet of Things (IoT) devices has led societies to new methods and searches.

IoT is an acronym that stands for "Internet of Things" and was first used in a presentation by Kevin Ashton in 1999. This concept covers systems that can exchange data through technological devices without human influence [14]. With the increase in the number of devices on the network, a serious increase in data traffic and security risks arise. The number of Internet-connected devices is expected to exceed 70 billion by 2025 [15].

The main goal of IoT is to enable objects to communicate with each other via the internet. In this context, it is anticipated that IoT can be integrated with almost all objects. These objects use embedded systems to communicate with internal servers and the external environment. The term "thing" is a concept that refers to devices connected to the Internet and to each other. Some important objects are: [16].

Sensors: Sensors are devices that convert environmental physical features into electrical signals. In this way, they become processable by computers.

Controllers: Controllers, which convert the measurement data they receive from sensors into signals, are important intermediate devices that then transmit this data to the main devices. Controllers can transmit this data to devices or actuators in the cloud, an example of M2M (Machine To Machine) communication [17].

M2M Communication: The controllers' job is to collect data from sensors and provide an internet connection. At the same time, controllers can make instant decisions and transmit data to more powerful computers for analysis. This means that controllers can communicate with powerful computers located on the same network, as well as access powerful computers in remote locations via an internet connection. Data is transmitted to data centers over the internet via routers. This interaction is as shown in Fig. 1 [17].

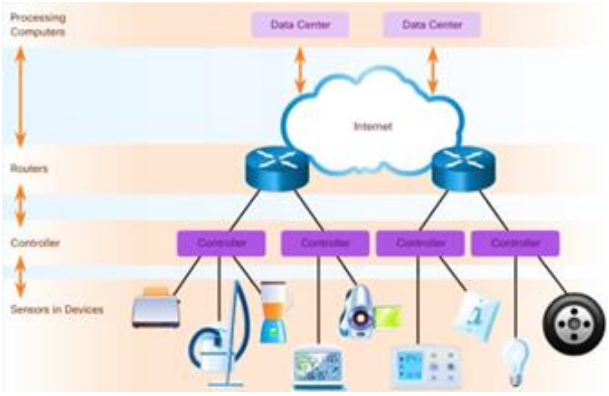


Fig.1. M2M Communication

Thanks to IoT, the amount of data will increase even more as devices that are not connected to the internet are integrated into the internet. The analysis of this large amount of data in the digital world is called Big Data. Big data is a new concept that describes heterogeneous data in various volumes that cannot be processed by traditional database techniques. This concept consists of different digital contents [18]. This data is obtained from online transactions, emails, videos, audio files, log records, search queries, health records, social network interactions, scientific data, sensors, mobile devices and applications [19,20].

There are three key features to consider for big data. These are:

Volume: It refers to the amount of data moved and stored. Nowadays, the amount of data is unpredictably large and constantly increasing. For example, 10 billion messages are sent per day on Facebook alone. Sensors, machines, cameras, etc. Mobile cameras, which are recording at all times, constantly produce data and expand the data volume [21].

Variety: Indicates the type of data. Data can occur in three different types: structured, semi-structured and unstructured. Indeed, most of the data produced today is of unstructured type [22].

Velocity: Data constantly changes dynamically. The production speed of big data is quite high and this speed is increasing day by day. From this perspective, it is important that the processes that will process and analyze the data keep up with the production speed of big data. Today, data must be produced very quickly, disseminated very quickly and analyzed very quickly [23].

The main purpose of big data management is to discover the data value hidden in this big data. The above-mentioned features of big data are used to achieve value. Value is seeing patterns, insights, relationships in data, discovering information from data, and predicting the future. To achieve these goals, effective data analysis must be performed. In this context, the purpose of big data is to collect data and transform it into important information [24].

IV. The Relationship Between the Internet of Things and Social Media Users

The relationship of data obtained from social platforms with the Internet of Things (IoT) may vary depending on various scenarios and use cases. But in general, there can be several points of connection between IoT and social platform data [25].

A. Features of Social Media Users

Characteristics of major social media users play an important role in analyzing their interactions on the platform. These features can be determined through various metrics and used to understand user behavior, define the target audience, and measure interactions on the platform. Features that focus on specific metrics for social media users include: [26].

Total Number of Posts: The total number of posts shared by the user on the platform reflects the user's activity level and is an important metric for measuring social media activity [27].

User Registration Date: It is a feature that shows when the user joined the social media platform, the user's experience and commitment to the platform.

Following and Number of Followers: The number of people the user follows and the number of people who follow him/her are important metrics that reflect the user's interaction on the social network and the size of the network.

Location Information: Whether the user shares location information in their posts is important for understanding geographic and local interactions.

Number of Likes: The number of likes received by the user's posts is an important metric that shows the popularity of their content and its impact on their followers.

These features provide essential data for social media analytics and offer valuable insights for businesses, marketers, or researchers to understand user behavior and improve platform strategies. These metrics, especially when integrated with IoT, guide us with the big data obtained in determining the accuracy and reliability of the target audience [28].

B. Architectural Connection of Social Media Users and IoT

Data Collection and Analysis (Sensor Layer): Various actions are performed depending on the data detected in the physical environment. Different types of data are retrieved from the real world by various sensors. This layer consists of multiple sensors. Social media data includes information such as users' various interactions and posts. This data is similar to the physical world data collected by the IoT sensor layer [29].

Data Communication (Network Layer): IoT systems have a network layer that enables data communication between devices. Social media platforms are also built on a network that facilitates communication between users. Therefore, digital interactions between social media users may share similar functionality of this layer [30].

Data Storage and Processing (Application and Analytics Layer): This layer performs tasks such as processing, analyzing, evaluating and making decisions based on the information it receives from the lower layer, and also has the responsibility of sharing this information with other devices. IoT systems typically store, analyze and process data. Social media data also contains a large amount of user interaction, and this data is often analyzed and used in various applications. Analyzes on user behavior can focus specifically on the application and analytics functionality of this layer [31].

Decision Making and Control (Application Layer): IoT systems are generally systems that can make decisions based on certain conditions and can control devices. Social media data can help make decisions based on specific user profiles, such as ad targeting, and drive content recommendations. Fig. 2 shows the layers that make up traditional IoT systems [32].

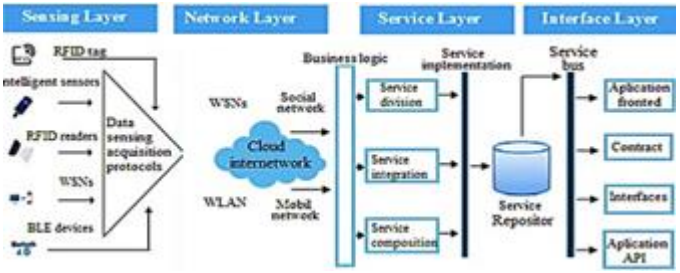


Fig.2. Traditional IoT Layers

The Sensing Layer includes sensing devices such as smart sensors, radio frequency identification (RFID) and client components of IoT to detect and obtain information [33].Network Layer is the layer that supports the connection infrastructure with the internet and other devices [27].Service Layer is the layer where providing and managing services to users or other applications takes place [34].Application Layer (Interface Layer) provides interface to users or other services [35].

V. METHODOLOGY

The relationship of data obtained from social platforms with the Internet of Things (IoT) may vary depending on various scenarios and use cases. But in general, there can be several points of connection between IoT and social platform data. This research aims to examine the interactions and accuracy of social media users in detail, using a comprehensive data set I collected from social media platforms. The data set used in the study includes information of a total of 17,500 social media users. 6 different features and a class of tag information in the profiles of these users form the basis of this research.

A. Data Collection and Preprocessing

The social media data used in this study was obtained through a manual data collection process using the Selenium automation tool. Selenium enables programmatic control of web browsers, allowing for the automatic extraction of user profiles and interaction data from social media platforms. The collected dataset includes various features of 17,500 social media users, and data integrity and consistency were carefully maintained throughout the data collection process.

Following the data collection phase, a comprehensive data preprocessing stage was carried out to enhance the accuracy of the analyses. The preprocessing involved the following steps:

Missing Data Analysis and Imputation: Missing or null values in the dataset were identified. These values were either filled using mean/median imputation or excluded from the analysis if they contained excessive missing information.

Data Cleaning: Duplicate records and meaningless values (e.g., data containing logical inconsistencies) were removed from the dataset.

Feature Engineering: Categorical variables were converted into numerical format using methods such as label encoding or one-hot encoding to make them suitable for machine learning algorithms.

Normalization and Standardization: Due to the different scales of numerical variables, Min-Max normalization or Z-score standardization was applied to ensure better learning performance of the algorithms.

Dataset Splitting: The dataset was split into training and test sets, typically at a ratio of 70% training and 30% testing. Care was taken to ensure a balanced distribution of data across both subsets.

B. Model Training and Hyperparameter Tuning

During the model training phase, the processed dataset was divided into 70% training and 30% testing subsets to objectively evaluate the performance of the machine learning algorithms. While models were trained on the training data, the test data was reserved to assess the models' generalizability and accuracy.

The algorithms employed in this study include Logistic Regression, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Random Forest. These models were selected to predict the authenticity status of social media users.

To optimize the performance of each model, hyperparameter tuning was conducted. Hyperparameter optimization is a crucial step for enhancing model accuracy and stability and was implemented through the following techniques:

Grid Search: All combinations within the predefined hyperparameter ranges were tested, and the parameter set yielding the highest accuracy score was selected.

Cross-Validation: To prevent overfitting, 5-fold cross-validation was applied. The training data was divided into multiple subsets to test the model's consistency across different data splits.

Examples of tuned hyperparameters include:

Table II. Hyperparameters tuned for each machine learning algorithm.

Model	Tuned Hyperparameters
Logistic Regression	Regularization strength (C), maximum iterations
k-NN	Number of neighbors (k), distance metric (e.g., Euclidean)
SVM	Kernel type (linear, RBF), C, gamma
Random Forest	Number of trees (n_estimators), max depth, min samples per leaf

Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC (Area Under the ROC Curve). These metrics were selected to provide a balanced view of classification performance, accounting for both general and class-specific performance.

After completing training with the best-performing hyperparameter configurations, the models were evaluated on the test data and the results were compared. This process helped identify which algorithm most effectively detected the accuracy of IoT-derived social media data.

The metrics in the user profiles examined vary widely and include important features such as user interactions, shares, follower-following numbers, account opening year, and location. These metrics are evaluated to understand the behavioral characteristics of social media users and to determine the accuracy of IoT-derived big data based on these characteristics. Additionally, a correlation matrix was created to understand the relationships between the features used. The correlation matrix will help us better understand the performance of fact detection algorithms by identifying strong or weak relationships between social media metrics.

In the analysis phase, leading machine learning algorithms such as Logistic Regression, k-NN, Support Vector Machines and Random Forest were applied, using 30% of the learning data. The remaining 70% test data was used to evaluate the performance of these algorithms. The features used in the study were carefully selected to ensure that the learning algorithms take into account features that may or may not directly affect the

accuracy of social media users. This method increased the reliability of the analyzes in order to obtain accurate and meaningful results.

As a result, this research focuses on comparing different machine learning algorithms to detect the accuracy of IoT-derived big data on social media. Detailed analyzes reveal the effectiveness of social media data in determining the accuracy of social media data by evaluating the performance of the algorithms used in detail. This comprehensive review makes a significant contribution to how social media and IoT integration can be optimized for accuracy. The correlation matrix is presented in Fig. 3.

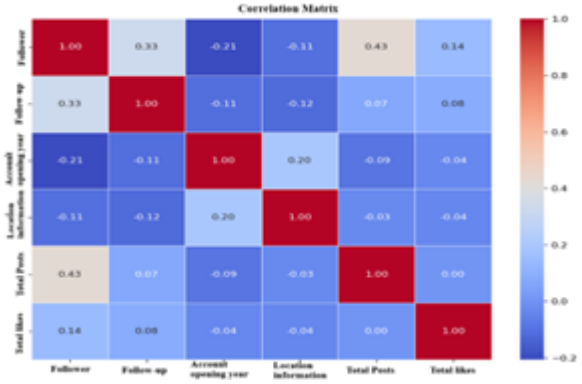


Fig.3. Correlation Matrix

The correlation matrix presented in Fig. 3 is interpreted as follows. Follower and Following Relationship: There is a positive correlation of 0.33 between the number of followers and the number of people followed. This shows that, in general, as a person's number of followers increases, the number of people that person follows tends to increase as well. Relationship between Followers and Account Opening Year: There is a negative correlation of -0.21 between the number of followers and the year the account was opened. This shows that newer users generally have fewer followers. Relationship between Account Opening Year and Location Information: There is a positive correlation of 0.20 between account opening year and location information. This suggests that, in general, newer accounts are more likely to share location information. Relationship between Location Information and Fake Account: A perfect positive correlation of 1.00 between location information and fake account is noteworthy. However, this probably indicates an error or bias in the data set, as such a high correlation often cannot be explained in a logical context. Total Likes and Fake Account Relationship: There is a negative correlation of -0.04 between the total number of likes and fake accounts. This shows that the total number of likes does not correlate much with the creation of fake accounts. These interpretations are based on the magnitude and sign of the values in the correlation matrix. However, it should be noted that correlation does not indicate a causal relationship, it only shows the strength of the relationship between two variables and how directionally related they are.

VI. Experimental Results

A. Classification Reports

In this research, various performance criteria were used to evaluate the findings. Metrics such as accuracy, precision, recall, and F1-score are used to evaluate the effectiveness of the classification model, so it is possible to understand how successful the model is and make improvements [36].

An attempt was made to determine the success of the classification algorithms used. Accuracy is an important metric used to evaluate the overall performance of a model. In physical property measurement, it is the discrepancy between the actual value and the value measured by the device [37]. A high precision rate indicates that the model classifies correctly, whereas a low precision rate indicates that the model classifies incorrectly. This criterion is used when evaluating the overall effectiveness of the model and provides information on the reliability of the model's performance [38]. The accuracy estimate is represented by Eq. (1).

$$Acccuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

Success (Precision) is a clear indication of the expected situation. This is an indication of the age of the true positives claimed by a percentage model in relation to all positives. It is used to make a precise determination regarding the demands of the model. It is represented by Eq. (2)

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall Rate focuses on total positives. The system refers to a certain number of definite positives in the data. It is represented by Eq. (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

The F1 score can be used to evaluate model performance. It is the harmonic mean of recall and precision values. It is represented by Eq. (4).

$$F1\ Score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

The metrics of the data used for validation and prediction and the effect of the Logistic Regression approach on classification are presented in Table 3.

Table III. Logistic regression classification.

	Precision	Recall	F1-Score	Support
0	0.61	0.25	0.35	661
1	0.55	0.85	0.67	712
Accuracy	-	-	0.51	1373
Macro Avg.	0.58	0.55	0.51	1373
Weight Avg.	0.58	0.56	0.52	1373

The metrics of the data used for verification and prediction and the effect of the Naive Bayes approach on classification are presented in Table 4.

Table IV. Naive bayes classification report.

	Precision	Recall	F1-Score	Support
0	0.61	0.08	0.14	661
1	0.52	0.95	0.67	712
Accuracy	-	-	0.41	1373
Macro Avg.	0.57	0.51	0.41	1373
Weight Avg.	0.57	0.53	0.42	1373

The metrics of the data used for verification and prediction and the impact of the Random Forest approach on classification are presented in Table 5. classification report.

Table V. Random forest classification report.

	Precision	Recall	F1-Score	Support
0	0.56	0.59	0.57	661
1	0.60	0.57	0.58	712
Accuracy	-	-	0.58	1373
Macro Avg.	0.58	0.58	0.58	1373
Weight Avg.	0.58	0.58	0.58	1373

The metrics of the data used for verification and prediction and the effect of the K-nearest neighbor approach on classification are presented in Table 6.

Table VI. K-nearest neighbor classification report.

	Precision	Recall	F1-Score	Support
0	0.54	0.52	0.53	661
1	0.57	0.58	0.57	712
Accuracy	-	-	0.55	1373
Macro Avg.	0.55	0.55	0.55	1373
Weight Avg.	0.55	0.55	0.55	1373

The metrics of the data used for verification and prediction and the effect of the Support Vector Machines approach on classification are presented in Table 7.

Table VII. Support Vector Machines classification report.

	Precision	Recall	F1-Score	Support
0	0.61	0.42	0.49	661
1	0.58	0.75	0.65	712
Accuracy	-	-	0.57	1373
Macro Avg.	0.59	0.58	0.57	1373
Weight Avg.	0.59	0.59	0.58	1373

Table 8 shows the accuracy values of the classification algorithms.

Table VIII. Accuracy scores report.

Algorithms	Accuracy Scores
Logistic Regression	0.51
Naive Bayes	0.41
Random Forest	0.58
K-Nearest Neighbor	0.55
Support Vector Machines	0.57

Table 9 shows the data processing speeds of Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), K-nearest neighbor and Support Vector Machines (SVM) algorithms depending on their execution time in seconds.

Table IX. Execution time performance report.

Algorithms	Execution Time(sec)
Logistic Regression	0.01
Naive Bayes	0.02
Random Forest	0.32
K-Nearest Neighbor	0.01
Support Vector Machines	0.03

As shown in Table VIII, among the evaluated algorithms, **Random Forest achieved the highest accuracy (0.58), followed closely by Support Vector Machines (0.57) and K-Nearest Neighbor (0.55),** while Logistic Regression (0.51) and Naive Bayes (0.41) performed comparatively lower. This indicates that ensemble and margin-based algorithms are more effective in detecting accuracy and authenticity in the given SIoT dataset.

Table IX highlights the **execution times** of each algorithm. Logistic Regression and K-Nearest Neighbor exhibited the fastest processing times (0.01 sec), whereas Random Forest required the longest time (0.32 sec) due to its ensemble nature. These results demonstrate a **trade-off between model accuracy and computational cost**, suggesting that the choice of algorithm should consider both performance and efficiency depending on the application scenario.

Overall, the comparative analysis provides insights into the **strengths and limitations of different ML models**, guiding the selection of the most suitable method for large-scale SIoT data accuracy and authenticity detection.

B. Performance Curve

A graph known as a ROC curve shows the performance of a classification model [39].It is defined as the area under the ROC curve (AUC). AUC gives an average performance value that summarizes the ROC curve. As the ROC curve approaches the upper left corner, it shows that the overall accuracy of the test increases, so it is preferable for the AUC value to be close to one [40].

Logistic Regression: The size of the area under the ROC curve of the logistic regression model is 0.63. This indicates that the model's ability to distinguish positive and negative classes is moderate.

Naive Bayes: The size of the area under the ROC curve of the Naive Bayes model is 0.58. This indicates that the model performs slightly lower than other models.

Random Forest: The size of the area under the ROC curve of the Random Forest model is 0.62. This indicates that the model's ability to distinguish positive and negative classes is moderate.

K-Nearest Neighbor: The size of the area under the ROC curve of the K-nearest neighbor model is 0.58. This indicates that the model performs slightly lower than other models.

Support Vector Machines (SVM): The size of the area under the ROC curve of the SVM model is 0.64. This shows that SVM performs better than other models.

C. Confusion Matrix

A criterion used to classify algorithm results is represented by the confusion matrix. The columns of this matrix represent the true values of the samples, whereas the rows correspond to the classification results. In this research, criteria such as sensitivity, selectivity, F-measure, accuracy, balanced accuracy and area under the ROC curve were used when evaluating classification success. The confusion matrix includes predicted classes and actual classes as a result of classification of labeled data [41].

True Positive: Reviews correctly classified as positive.

True Negative: Comments correctly classified as negative.

False Positive: Reviews incorrectly classified as positive.

False Negative: Reviews incorrectly classified as negative.

Fig. 4 presents the confusion matrix for the Logistic Regression algorithm.

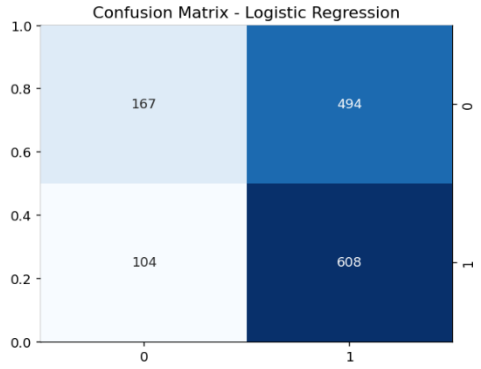


Fig.4. Logistic Regression Confusion Matrix

True Positive: This indicates that 167 samples were classified correctly.

False Positive: This case, 494 instances were misclassified as class 0.

False Negative: The actual class is 1 but the predicted class is 0. In this case, 104 samples were misclassified as class 1.

True Negative: The true class is 1 and the predicted class is 1. Here, 608 examples were correctly classified.

This matrix shows that the Logistic Regression classification model makes some correct predictions, but also makes some incorrect classifications.

Fig. 5 presents the confusion matrix for the Naive Bayes algorithm.

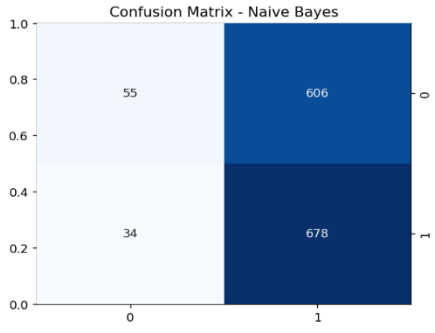


Fig.5. Naive Bayes Confusion Matrix

True Positive: Here it is shown that 55 samples were correctly classified.

False Positive: This case has 606 instances misclassified as class 0.

False Negative: 34 samples were misclassified as class 1.

True Negative: 678 samples were correctly classified.

This matrix shows that the model identifies class 0 effectively but describes class 1 more poorly. Whereas the ability to correctly classify samples belonging to class 0 is higher, both false positive (FP) and false negative (FN) values for class 1 are proportionally high, indicating that the model has a harder time identifying this class.

Fig. 6 presents the confusion matrix for the Random Forest algorithm.

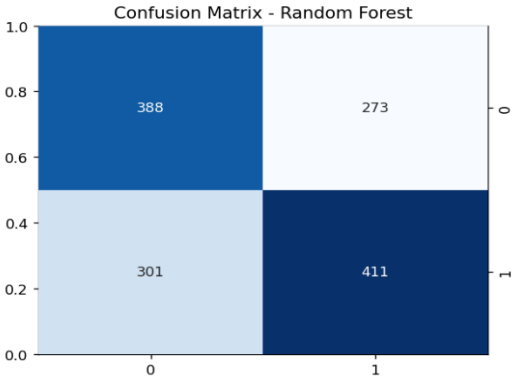


Fig.6. Random Forest Confusion Matrix

True Positive: 398 samples were classified correctly. The model correctly identified class 1

False Positive: 276 samples were misclassified as class 0.

False Negative: 314 samples were misclassified as class1.

True Negative: 385 samples were correctly classified. The model correctly identified class 0.

Whereas this matrix was successful in correctly identifying class 1, the relatively high number of both false positive and false negative values for class 0 indicates that it is poor at identifying class 0.

Fig. 7 presents the confusion matrix for the K-nearest neighbor algorithm.

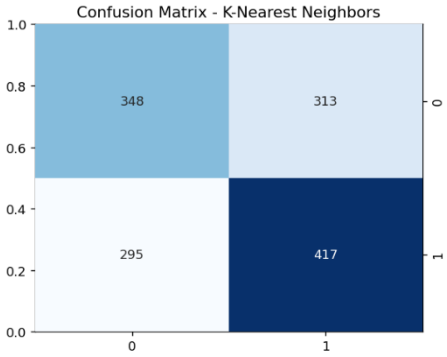


Fig.7. K-Nearest Confusion Matrix

True Positive: 417 samples were correctly classified.

False Positive: 313 samples were misclassified as class 0.

False Negative: 295 samples were misclassified as class 1.

True Negative: 348 samples were correctly classified. The model correctly identified class 0.

This matrix shows that the model is successful in identifying class 1. However, it refers to a situation where the false positive and false negative values are relatively high, indicating that it fails to identify class 0.

Fig. 8 presents the confusion matrix for the Support Vector Machines algorithm.

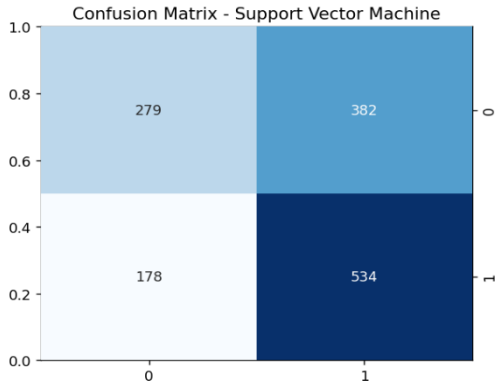


Fig.7. Support Vector Machines Confusion Matrix

True Positive: 534 samples were correctly classified.

False Positive: 382 samples were misclassified as class 0.

False Negative: 178 samples were misclassified as class 1.

True Negative: 279 samples were correctly classified.

This matrix shows that the model is successful in identifying class 1 but fails in identifying class 0.

VII. CONCLUSIONS

This study thoroughly evaluated the accuracy and reliability of multiple machine learning classification algorithms—Logistic Regression, Naive Bayes, Random Forest, K-Nearest Neighbor (K-NN), and Support Vector Machines (SVM)—using key performance metrics including precision, recall, F1-score, accuracy, and area under the ROC curve (AUC). Among these, Random Forest and SVM exhibited superior predictive performance, achieving accuracy scores of 0.58 and 0.57 respectively, with SVM attaining the highest AUC of 0.64. The ensemble learning strategy of Random Forest and the margin-maximizing optimization of SVM contributed to their effectiveness in managing the high-dimensional and noisy features characteristic of IoT-derived social media data.

Hyperparameter tuning was rigorously conducted via grid search and 5-fold cross-validation to optimize model generalizability and prevent overfitting. Parameters such as regularization strength for Logistic Regression, number of neighbors and distance metrics for K-NN, kernel type and gamma for SVM, and tree count and depth for Random Forest were systematically optimized.

Correlation analysis revealed expected positive relationships among social features, like follower and following counts, consistent with typical user behavior patterns. However, an unexpected perfect positive correlation between location data and fake account labels indicates potential data bias or labeling noise, underscoring the necessity for careful data validation and possible feature re-engineering in future studies.

While SVM demonstrated high true positive identification, it struggled with false positives, whereas Naive Bayes showed limited discriminatory power, likely due to its strong independence assumptions not fully aligning with the dataset's characteristics.

Despite these encouraging findings, several limitations exist. The dataset of 17,500 users, although sizable, may not fully capture the diversity of social media behaviors, limiting model generalizability. The manual data collection process introduces potential sampling biases and inconsistencies. Additionally, some features exhibited noise or bias that may affect model performance. Importantly, this study focuses solely on traditional machine learning algorithms, leaving the exploration of advanced deep learning architectures, which might better capture complex temporal and relational data patterns, for future work.

Given the increasing spread of misinformation amplified by AI on social networks, future research should develop integrated, AI-driven anomaly detection systems that enhance the identification and prevention of manipulated or false information. Specifically, the application of deep learning models such as recurrent neural networks and graph neural networks, combined with automated data cleaning and bias mitigation techniques, promises substantial improvements in model accuracy and fairness.

Furthermore, expanding the dataset to include more users and multiple social media platforms will improve representativeness and robustness. Enriching the feature space with temporal activity metrics, network centrality measures, and textual content analysis can enhance model expressiveness and practical relevance.

In conclusion, this study contributes foundational insights into applying machine learning to assess the accuracy and authenticity of IoT-derived social media data. Addressing current limitations and integrating advanced methodologies will drive future efforts towards more reliable, trustworthy SIoT data analytics frameworks, ultimately fostering safer and more transparent digital ecosystems.

VIII. FUTURE WORK

Despite the promising results of this study, several limitations should be noted. First, the dataset is limited to specific SIoT sources, which may restrict the generalizability of the findings to other domains such as healthcare, transportation, or industrial IoT. Second, while multiple machine learning algorithms were evaluated, more advanced deep learning approaches, including Graph Neural Networks or Transformer-based models, were not considered and could potentially improve accuracy and authenticity detection. Third, the current study focuses on static datasets; real-time streaming SIoT data scenarios were not addressed, which may affect the scalability and responsiveness of the proposed models. Additionally, feature selection and engineering were primarily based on existing attributes, leaving room for the integration of temporal, contextual, or semantic features to further enhance model performance.

For future work, we plan to expand the methodology to diverse SIoT domains and incorporate advanced ML and deep learning techniques. Investigating real-time data streams, developing standardized large-scale benchmark datasets, and integrating authenticity detection with trustworthiness or anomaly detection frameworks are also important directions. These steps aim to enhance the applicability, robustness, and reliability of machine learning-based accuracy and authenticity detection in SIoT environments.

REFERENCES

[1] Shahab, S., Agarwal, P., Mufti, T., & Obaid, A. J. (2022). SIoT (Social Internet of Things): A Review. *Evolutionary Computing and Mobile Sustainable Networks*, 313–323.

[2] Dhelim, S., Ning, H., Farha, F., Chen, L., Atzori, L., & Daneshmand, M. (2021). IoT-Enabled Social Relationships Meet Artificial Social Intelligence. *IEEE Internet of Things Journal*, 8(20), 15364–15375.

[3] Nejad, H. V., Farimani, Z. M., & Tavakolifar, A. (2020). Social Internet of Things and New Generation Computing—A Survey. *Toward Social Internet of Things (SIoT)*, 846, 129–152.

[4] Kaya, Ş. M., & Kaya, E. (2022). The (Un)seen Influence of S-IoT on the Political Economic Decisions. In *6th International Congress of Social Sciences*, Istanbul.

[5] Rad, M. M., Rahmani, A. M., Sahafi, A., & Qader, N. N. (2020). Social Internet of Things: vision, challenges, and trends. *Human-centric Computing and Information Sciences*, 10.

[6] Kaur, N., & Sood, S. K. (2023). Social Internet of Things (SIoT): A decade's journey and future directions. *Journal of Network and Computer Applications*, 210.

[7] Dhelim, S., Ning, H., Farha, F., Chen, L., Atzori, L., & Daneshmand, M. (2021). IoT-Enabled Social Relationships Meet Artificial Social Intelligence. *IEEE Internet of Things Journal*, 8(20), 15364–15375.

[8] İşler, B., Kaya, Ş. M., & Kılıç, F. R. (2025). Fog-Enabled Machine Learning Approaches for Weather Prediction in IoT Systems: A Case Study. *Sensors*, 25(13), 4070.

[9] Zhang, L., et al. (2021). Integrating Social Media Data with IoT for Enhanced User Experience. *IEEE Internet of Things Journal*, 8(3), 1540–1552.

[10] Kim, J., et al. (2020). Understanding User Characteristics and Interactions on Social Media Platforms. *Journal of Interactive Advertising*, 20(3).

[11] Chen, M., et al. (2019). Trust management in social Internet of Things: A survey. *IEEE Communications Surveys & Tutorials*, 22(2), 1197–1230.

[12] Li, J., & Wang, H. (2020). A sentiment-aware framework for social IoT applications based on hybrid machine learning. *Future Generation Computer Systems*, 108, 512–524.

[13] Gupta, R., et al. (2022). Real-time analytics for social IoT using edge computing. *IEEE Transactions on Network and Service Management*, 19(1), 67–80.

[14] Zhou, J., Leung, V. C., & Yang, L. T. (2021). Internet of Things security and privacy: Challenges and solutions. *IEEE Internet of Things Journal*, 8(12), 10231–10255.

[15] Kaya, Ş. M. (2025). Edge And Fog Computing With Artificial Intelligence Methods On Iot-Based Big Data. *Artificial Intelligence: Foundations, Applications and Future Directions*, 347.

[16] Johnson, E. (2019). IoT Sensors and Their Applications in Smart Systems. *Sensors and Actuators B: Chemical*, 185, 230–245.

[17] Hatton, M. (2013, January). The global M2M market in 2013. *Machina Research White Paper*.

[18] Gahi, Y., Guennoun, M., & Mouftah, H. T. (2016). Big Data Analytics: Security and Privacy Challenges. In *Proceedings of the 2016 IEEE Symposium on Computers and Communication (ISCC)* (pp. 952–957). Messina, Italy.

[19] Zikopoulos, I., Eaton, C. P., & Zikopoulos, P. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (1st ed.). McGraw-Hill Osborne Media.

[20] Schneider, R. D. (2012). *Hadoop for Dummies* (Special ed.). John Wiley & Sons.

[21] Setty, K., & Bakhshi, R. (2013). What Is Big Data and What Does It Have to Do with IT Audit?. *ISACA Journal*, 3, 23–25.

[22] Kaya, Ş. M., Bayram, V., & Özkan, M. (2025). Evaluation of the intergenerational relationship of IoT awareness in businesses. *Journal of Information and Optimization Sciences*, 46(5), 1753–1772.

[23] Cyganek, B., et al. (2016). A Survey of Big Data Issues in Electronic Health Record Analysis. *Applied Artificial Intelligence*, 30(6), 497–520.

[24] Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.

[25] Zhang, L., et al. (2021). Integrating Social Media Data with IoT for Enhanced User Experience. *IEEE Internet of Things Journal*, 8(3), 450–465.

[26] Esmaili Jobani, A., & Kaya, Ş. M. (2025). Hybrid IoT and AI-based Solution for Energy Management in Data Centres under Various Climate Conditions. *Anadolu Bil Meslek Yüksekokulu Dergisi*, 20(72), 107–124.

[27] Chen, Y., et al. (2019). Analyzing User Behavior on Social Media Platforms: Methods and Applications. *ACM Transactions on Social Computing*, 4(2), 75–90.

[28] Smith, E., et al. (2018). User Characteristics and Interactions on Social Media Platforms: Insights from Data Analytics. *International Journal of Information Management*, 45, 210–225.

[29] Hancke, G. P., & Hancke Jr., G. P. (2013). The role of advanced sensing in smart cities. *Sensors*, 13(1), 393–425.

[30] Talari, S., et al. (2017). A Review of Smart Cities Based on the Internet of Things Concept. *Energies*, 10(4), 421.

[31] Sikder, A. K., Petracca, G., Aksu, H., Jaeger, T., & Uluagac, A. S. (2018). A Survey on Sensor-Based Threats to Internet-of-Things (IoT) Devices and Applications. *ArXiv Preprint.*

[32] Kaya, Ş. M., Erdem, A., & Güneş, A. (2021). A Smart Data Pre-Processing Approach to Effective Management of Big Health Data in IoT Edge. *Smart Homecare Technology and TeleHealth*, 9–21.

[33] Kaya, Ş. M., İşler, B., Abu-Mahfouz, A. M., Rasheed, J., & AlShammari, A. (2023). An Intelligent Anomaly Detection Approach for Accurate and Reliable Weather Forecasting at IoT Edges: A Case Study. *Sensors*, 23(5), 2426.

[34] Ahmed, I., Saeed, A., & Malik, H. (2023). A trust-aware data filtering framework for Social Internet of Things. *Computer Networks*, 225, 109554.

[35] Kaya, Ş. M., Erdem, A., & Güneş, A. (2022). Anomaly Detection and Performance Analysis by Using Big Data Filtering Techniques For Healthcare on IoT Edges. *Sakarya University Journal of Science*, 26(1), 1–13.

[36] Li, S., Raymond, K. K., Sun, Q., Buchanan, W. J., & Cao, J. (2015). IoT Forensics: Amazon Echo as a Use Case. *Journal of Latex Class Files*, 14.

[37] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd ed.). Springer.

[38] Bayram, V., & Kaya, M. (2023). The Contributions of Metaverse Technology on Management Information Systems in Strategic Planning and Decision-Making Processes of Businesses. *Uluslararası Yönetim Akademisi Dergisi*, 6(3), 794–807.

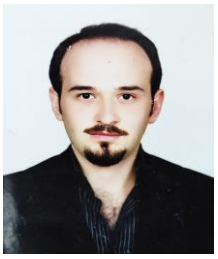
[39] Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192

[40] Kaya, Ş. M., & Bayram, V. (2025). Artificial Intelligence Awareness Scale Development Study. *OPUS Journal of Society Research*, 22(4), 657–672.

[41] Cihan, P. (2018). Determination of diagnosis, prognosis and risk factors in animal diseases using by data mining methods. *PhD Thesis*, Yildiz Technical University, Istanbul, Turkey.

BIOGRAPHIES

Deniz Kızılaslan received his BS degree in Computer Engineering from Sakarya University in 2016 and his MS degree in Computer Engineering from Istanbul Aydın University in 2025. His graduate research focused on Sentiment Analysis and Content Prediction using Machine Learning and Deep Learning methodologies. He has 10 years of professional experience in the software industry, currently serving as a Senior Software Developer. His professional expertise and current research interests include Generative AI, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and Natural Language Processing (NLP). He is actively involved in developing advanced software solutions and conducting research on data authenticity and predictive modeling in large-scale data.



Assist. Prof. Dr. Şükrü Mustafa Kaya graduated from Istanbul Aydin University, Institute of Graduate Education, Department of Computer Engineering in 2021 with his doctoral thesis "Smart Data Preprocessing Approach For Effective Management Of Health Big Data In IoT Edge". He started working as an Assistant Professor at Istanbul Aydin University, Department of Computer Technologies in 2022, took office as the Director of the Blockchain Application Research Center in November 2024 and continues to conduct academic studies within the scope of IoT as an Assistant Professor at Istanbul Aydin University, Department of Computer Technologies.