Айжан Конурбаевна Токкулиева – техника ғылымдарының магистрі, «Ақпараттық қауіпсіздік жүйелері» мамандығының 1 курс докторанты, Л.Н. Гумилёв атындағы Еуразия ұлттық университеті; «WebTotem» ЖШС кіші ғылыми қызметкері, Астана қ., Қазақстан; e-mail: aizhantokkuliyeva1983@gmail.com. ORCID: https://orcid.org/0000-0002-5019-2413.

Information about the authors

Aigul Kairulaevna Shaikhanova* – PhD, Professor of the Department of Information Security; L.N. Gumilyov Eurasian National University; Coordinator of scientific projects of WebTotem LLP, Astana, Kazakhstan; e-mail: aigul.shaikhanova@gmail.com. ORCID: https://orcid.org/0000-0001-6006-4813.

Ruslan Arimbaevich Budenov – Master of Technical Sciences, Electronics Engineer, WebTotem LLP, Astana, Kazakhstan; e-mail: akarkin@mail.ru. ORCID: https://orcid.org/0009-0003-9088-3221 .

Olzhas Shagzadovich Satiev – Senior Information security Expert at WebTotem LLP, Astana, Kazakhstan; e-mail: os@wtotem.com. ORCID: https://orcid.org/0009-0005-2684-1718.

Danir Amangeldinovich Tlepov – Master of Technical Sciences, Researcher at WebTotem LLP, Astana, Kazakhstan; e-mail: tdanir@cert.kz. ORCID: https://orcid.org/0009-0003-3774-8389.

Aizhan Konurbaevna Tokkuliyeva – Master of Technical Sciences, 1st year doctoral student, specialty «Information Security Systems», L.N. Gumilyov Eurasian National University; Junior Researcher at WebTotem LLP, Astana, Kazakhstan; e-mail: aizhantokkuliyeva1983@gmail.com. ORCID: https://orcid.org/0000-0002-5019-2413.

Редакцияға енуі 03.03.2025
Өңдеуден кейін түсуі 05.03.2025
Жариялауға қабылданды 12.03.2025

https://doi.org/10.53360/2788-7995-2025-1(17)-5

IRTSTI: 28.23.15

A.K. Kalpen[1], E.T. Matson[2], A.K. Zhumadillayeva[1,3*], K.A. Dyussekeyev[3]
[1]Astana IT University,
Kazakhstan, Astana, Mangilik El 55/11, Block C1 QazExpo
[2]Purdue University,
West Lafayette, Indiana, USA
[3]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
*e-mail: Zhumadillayeva_ak@enu.kz

SEQUENCE RECOGNITION USING FINITE AUTOMATA WITH MACHINE LEARNING

Annotation: *Sequence recognition is a critical task across numerous disciplines. While traditional methods utilizing Finite State Machines (FSMs) offer a structured data representation and high interpretability, their flexibility is limited. Contemporary Machine Learning (ML) algorithms exhibit high accuracy but demand substantial computational resources. Combining these paradigms can enhance the effectiveness of complex sequence recognition. This study explores the integration of FSMs with ML techniques to address sequence analysis problems. Three distinct applications are examined: text classification (spam detection), recognition of genetic sequences related to Alzheimer's disease, and image-based gesture identification.*

For each, hybrid models were developed and tested, combining Deterministic Finite Automata (DFA), Non-deterministic Finite Automata (NFA), and ML algorithms such as Random Forest, Gradient Boosting, and Multilayer Perceptrons (MLP). Experimental results indicate that these hybrid models achieve performance comparable to traditional ML methods, and in some instances, yield more accurate predictions.

In spam classification, neural network models demonstrated the best results, with FSM-neural network combinations providing similar effectiveness.

For genetic sequence analysis, gradient boosting-based models exhibited the highest accuracy, with the inclusion of FSMs maintaining performance while enhancing interpretability.

In gesture recognition, neural network approaches proved most effective, but integrating FSMs with ensemble methods achieved a high level of predictive capability, surpassing conventional ML models.

In conclusion, the integration of FSMs and ML presents a promising avenue in sequence analysis. Future research could focus on optimizing model architectures and applying them to other domains requiring high-precision recognition of intricate structures.

Key words: *Finite State Machine, Machine Learning, Sequence Recognition, Hybrid Models, Genetic Sequence Analysis, Gesture Recognition, Text Classification.*

Introduction

In the modern era of digital transformation and rapid data growth, sequence recognition tasks are becoming increasingly significant. The processing of text, genetic, and visual data necessitates the development of reliable and precise methods capable of uncovering hidden patterns within complex structures. Traditional approaches based on finite automata (FA) possess formal rigor and high interpretability; however, their application is limited by fixed structures and insufficient flexibility [1]. Specifically, deterministic finite automata (DFA) ensure predictability, while nondeterministic finite automata (NFA) offer enhanced flexibility, though both models encounter scalability issues when handling large datasets [1].

On the other hand, machine learning (ML) methods demonstrate remarkable adaptability and the ability to learn from vast amounts of data, which enables them to achieve high accuracy in recognizing intricate patterns [2]. However, ML models are often regarded as "black boxes", making their inner workings difficult to interpret, and they require considerable computational resources. In light of these characteristics, there is a pressing need to integrate FA and ML, combining the structured representation and formal precision of automata with the adaptability and high accuracy of modern machine learning algorithms [3].

The relevance of this research is driven by the growing demand for universal hybrid models for sequence recognition that can operate effectively under conditions of data variability and noise. Existing studies on the application of finite automata to tasks such as syntactic analysis, bioinformatics, and gesture recognition [4, 5] underscore the potential of a combined approach. At the same time, research focusing on the integration of ML methods with FA [6, 7] highlights the necessity for developing new algorithms capable of overcoming the limitations inherent in both traditional methods and current machine learning models.

The aim of the present study is to develop and experimentally evaluate hybrid models that integrate finite automata and machine learning methods to enhance the accuracy and interpretability of sequence recognition. In pursuit of this aim, the following tasks are addressed:

- Develop algorithmic approaches for integrating FA and ML that consider the characteristics of various data types.
- Conduct an experimental evaluation of the proposed hybrid models in tasks such as text classification, genetic sequence analysis, and gesture recognition.
- Identify the advantages and limitations of the integrated approach.

Methods

In this study, a hybrid approach is implemented that combines formal models of Finite State Machines (FSMs) with modern Machine Learning (ML) algorithms for sequence recognition. This approach enables the integration of the structured data representation provided by FSMs with the adaptability and high accuracy of ML models. The methodology is demonstrated using text data as an example.

1. **Data Preprocessing**

At the initial stage, comprehensive preprocessing of the raw data is performed: *Text Data.* Text normalization, conversion to lowercase, removal of extraneous characters, lemmatization, and stop-word filtering (while preserving key terms) are applied. For these purposes, the NLTK library and a custom class, *EnhancedTextPreprocessor*, are used.

```
class EnhancedTextPreprocessor:
    def __init__(self):
        self.lemmatizer = WordNetLemmatizer()
        self.stop_words = set(stopwords.words('english'))
        self.spam_keywords = {'free', 'win', 'prize', 'claim', 'urgent', 'offer', 'cash', 'limited'}

    def preprocess(self, text):
        text = str(text).lower()
        text = re.sub(r'[^\w\s\$\!\#]', '', text)  # Сохраняем спецсимволы (Save the wildcards)
        words = text.split()
        # Удаляем стоп-слова, кроме ключевых спам-слов (Remove stop words except for key spam words)
        words = [self.lemmatizer.lemmatize(word) for word in words
                 if word not in self.stop_words or word in self.spam_keywords]
        return ' '.join(words)
```

Figure 1 – EnhancedTextPreprocessor

Image and Numerical Data. Scaling (using StandardScaler) and normalization methods are applied to ensure that the features are correctly represented for further analysis.

2. **Implementation of Finite State Machines**

Two main models of FSMs are employed for modeling sequences: *Deterministic Finite Automata (DFA).* These provide predictability through fixed transitions between

states. The class *EnhancedDFA* implements this concept, enabling the processing of input sequences.

```
class EnhancedDFA:
    def __init__(self):
        self.states = {'q0', 'q1', 'q2', 'q3', 'q4', 'q5', 'q6'}
        self.alphabet = set('abcdefghijklmnopqrstuvwxyz$!1234567890')
        self.transitions = {
            'q0': {'s': 'q1', 'f': 'q3', 'w': 'q4', '$': 'q5', '!': 'q6', 'default': 'q0'},
            'q1': {'p': 'q2', 'default': 'q0'},
            'q2': {'a': 'q_accept', 'default': 'q0'},
            'q3': {'r': 'q_accept', 'default': 'q0'},
            'q4': {'i': 'q_accept', 'default': 'q0'},
            'q5': {'$': 'q_accept', 'default': 'q0'},
            'q6': {'!': 'q_accept', 'default': 'q0'},
            'q_accept': {'default': 'q_accept'}
        }
        self.start_state = 'q0'
        self.accept_states = {'q_accept'}

    def process_input(self, text):
        current_state = self.start_state
        for char in text:
            current_state = self.transitions[current_state].get(char,
                self.transitions[current_state].get('default', 'q0'))
            if current_state in self.accept_states:
                return 1
        return 0
```

Figure 2 – EnhancedDFA

Nondeterministic Finite Automata (NFA). These offer greater flexibility by allowing transitions to multiple states for a single input symbol. The class *EnhancedNFA* facilitates modeling of complex sequences, which is especially important when dealing with noisy data.

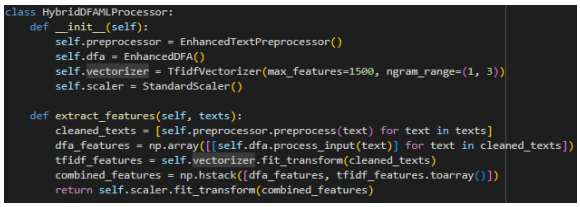
```
class EnhancedNFA:
    def __init__(self):
        self.states = {'q0', 'q1', 'q2', 'q3', 'q4', 'q5'}
        self.alphabet = set('abcdefghijklmnopqrstuvwxyz$!%1234567890')
        self.transitions = {
            'q0': {
                'f': {'q1'}, 'w': {'q2'},
                '$': {'q3'}, '!': {'q4'},
                'default': {'q0'}
            },
            'q1': {'r': {'q_accept'}, 'default': {'q0'}},
            'q2': {'i': {'q_accept'}, 'default': {'q0'}},
            'q3': {'$': {'q_accept'}, 'default': {'q0'}},
            'q4': {'!': {'q_accept'}, 'default': {'q0'}},
            'q_accept': {'default': {'q_accept'}}
        }
        self.start_states = {'q0'}
        self.accept_states = {'q_accept'}

    def process_input(self, text):
        current_states = self.start_states
        for char in text:
            next_states = set()
            for state in current_states:
                transitions = self.transitions[state].get(char,
                    self.transitions[state].get('default', set()))
                next_states.update(transitions)
            current_states = next_states
            if not current_states:
                break
            if 'q_accept' in current_states:
                return 1
        return 0
```

Figure 3 – EnhancedNFA

3. **Integration with Machine Learning Methods**

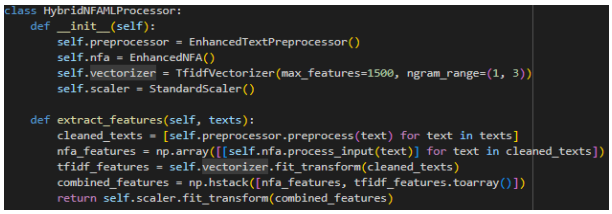
To improve the accuracy of sequence recognition, FSMs are integrated with ML algorithms: *Hybrid Models.* Classes such as *HybridDFAMLProcessor* and *HybridNFAMLProcessor* are used to form combined features. Initially, the raw data is processed by the FSMs (calculating binary features that reflect the passage of the sequence through the automaton) and then combined with features obtained via TF-IDF (for texts) or standard methods for processing numerical data.



```
class HybridDFAMLProcessor:
    def __init__(self):
        self.preprocessor = EnhancedTextPreprocessor()
        self.dfa = EnhancedDFA()
        self.vectorizer = TfidfVectorizer(max_features=1500, ngram_range=(1, 3))
        self.scaler = StandardScaler()

    def extract_features(self, texts):
        cleaned_texts = [self.preprocessor.preprocess(text) for text in texts]
        dfa_features = np.array([[self.dfa.process_input(text)] for text in cleaned_texts])
        tfidf_features = self.vectorizer.fit_transform(cleaned_texts)
        combined_features = np.hstack([dfa_features, tfidf_features.toarray()])
        return self.scaler.fit_transform(combined_features)
```

Figure 4 – HybridDFAMLProcessor



```
class HybridNFAMLProcessor:
    def __init__(self):
        self.preprocessor = EnhancedTextPreprocessor()
        self.nfa = EnhancedNFA()
        self.vectorizer = TfidfVectorizer(max_features=1500, ngram_range=(1, 3))
        self.scaler = StandardScaler()

    def extract_features(self, texts):
        cleaned_texts = [self.preprocessor.preprocess(text) for text in texts]
        nfa_features = np.array([[self.nfa.process_input(text)] for text in cleaned_texts])
        tfidf_features = self.vectorizer.fit_transform(cleaned_texts)
        combined_features = np.hstack([nfa_features, tfidf_features.toarray()])
        return self.scaler.fit_transform(combined_features)
```

Figure 5 – HybridNFAMLProcessor

ML Algorithms Employed. In the experimental part, ensemble methods (Random Forest, Gradient Boosting) and neural networks (MLP) are applied. The TensorFlow/Keras framework is used for implementing neural networks, while the scikit-learn library is used for ensemble methods.

4. **Experimental Setup**

The study includes experiments in three domains: *Text Classification (Spam Detection).* The SMS Spam Collection dataset is preprocessed, and hybrid features are generated using FSMs. Models are trained and evaluated using metrics such as accuracy and F1-score.

Genetic Sequence Analysis (Alzheimer's Disease). Additional features, generated based on simple rules implemented by FSMs, are combined with the original features. The same metrics are used for evaluation, allowing for comparison between hybrid and pure ML models.

Gesture Recognition. Data in the form of images are scaled and normalized. Hybrid models that combine FSMs with ensemble algorithms are applied for image classification, demonstrating high predictive capability.

5. **Tools and Technologies**

Programming Language**:** Python – for data processing and algorithm implementation.

Libraries:

Pandas, NumPy – for data manipulation;
NLTK, TfidfVectorizer – for text processing;
scikit-learn – for implementing and evaluating ML algorithms;
TensorFlow/Keras – for building neural networks;
Seaborn, Matplotlib – for result visualization.
Custom Classes: *EnhancedTextPreprocessor, EnhancedDFA, EnhancedNFA, HybridDFAMLProcessor, HybridNFAMLProcessor* – for integrating FSMs with ML methods.

Thus, the proposed methodology combines sequence analysis with adaptive machine learning algorithms, significantly enhancing the accuracy and interpretability of recognition systems.

Code link: https://colab.research.google.com/drive/1g5F16v7FtNLnyPxf7CgTxNB82ucwXHyA?usp=sharing

Results

Spam.

In this research, the SMS Spam Collection dataset was used for the task of text classification for spam detection.

Table 1 – Spam results

Model	Accuracy	F1-score
DFA+NN	0,982	0,930
NFA+NN	0,982	0,930
Neural Network	0,982	0,931
DFA+RF	0,979	0,917
NFA+RF	0,979	0,918
Random Forest	0,981	0,925
DFA+GB	0,970	0,877
NFA+GB	0,971	0,881
Gradient Boosting	0,971	0,882
Pure DFA	0,809	0,360
Pure NFA	0,809	0,360

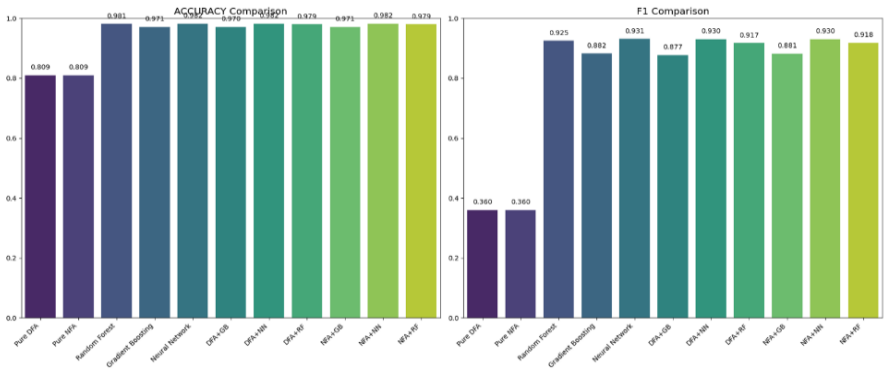


Figure 6 – Spam results

As can be seen from the table, pure finite automata are significantly inferior in terms of results, which emphasizes the need for their integration with adaptive machine learning algorithms, while hybrid models (e.g., DFA+NN and NFA+NN) in turn show the highest performance comparable to pure neural networks. This indicates that integrating finite automata with ML helps to improve recognition performance by generating additional binary features that reflect specific patterns in texts. However, despite the positive impact of such integration, the tuning of finite automata is critical and requires significant time and financial resources, which may make their use less feasible if more computing power is available.

DNA sequence.

In this experiment, the aim was to investigate the potential of hybrid models to analyze genetic sequences associated with Alzheimer's disease.

The study was conducted considering the characteristics of genetic data, where not only high recognition accuracy but also interpretability of the results obtained is important.

Table 2 – DNA results

Model	Accuracy	F1-score
DFA+NN	0,713132	0,659258
NFA+NN	0,715353	0,645664
Neural Network	0,716565	0,645031
DFA+RF	0,714815	0,640170
NFA+RF	0,717103	0,652616
Random Forest	0,715690	0,640572
DFA+GB	0,726392	0,675553
NFA+GB	0,726122	0,674662
Gradient Boosting	0,726392	0,675553
Pure DFA	0,606314	0,349605
Pure NFA	0,674766	0,685376

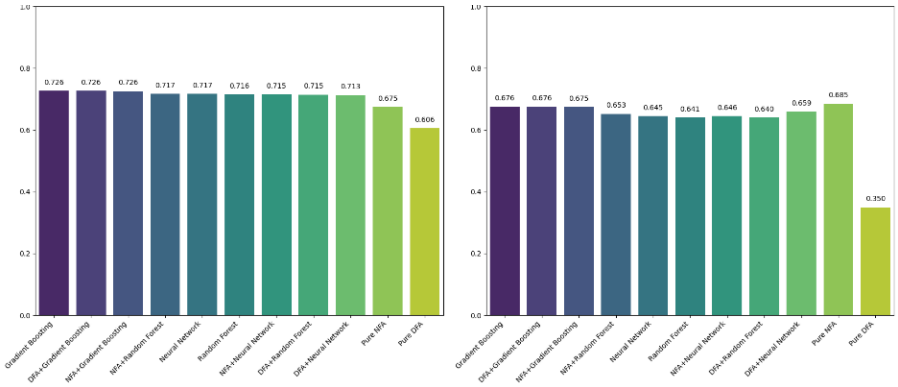


Figure 7 – DNA results

The table shows that hybrid models, in particular those using gradient boosting (DFA+GB, NFA+GB, and Pure Gradient Boosting), achieve the highest accuracy and F1-score. Meanwhile, models using only finite automata (Pure DFA) show significantly lower results, which confirms the insufficiency of their independent application in this task. The overall interpretability of the hybrid models is improved by explicitly highlighting significant patterns, which is important for bioinformatics and may contribute to a better understanding of the mechanisms underlying diseases. It is interesting to note that integration of finite automata not only improves the accuracy but also improves the interpretability of the model by explicitly highlighting important patterns in genetic sequences, but despite the positive effect of integration, tuning the finite automata requires additional time and resources. This can be a critical factor when scaling the system, as code development and optimization for hybrid models can be more costly compared to the increased computational power for pure ML models.

Sign Recognition.

The sign recognition task used images that were divided into pixels.

Table 3 – Sign Recognition results

Model	Accuracy	F1-score
DFA+NN	0,783324	0,782979
NFA+NN	0,780257	0,781183
Neural Network	0,799219	0,799182
DFA+RF	0,808979	0,810567
NFA+RF	0,665365	0,659743
Random Forest	0,667736	0,660938
DFA+GB	0,720998	0,726267
NFA+GB	0,486057	0,493412
Gradient Boosting	0,483547	0,491259
Pure DFA	0,041132	0,041901
Pure NFA	0,042387	0,043839

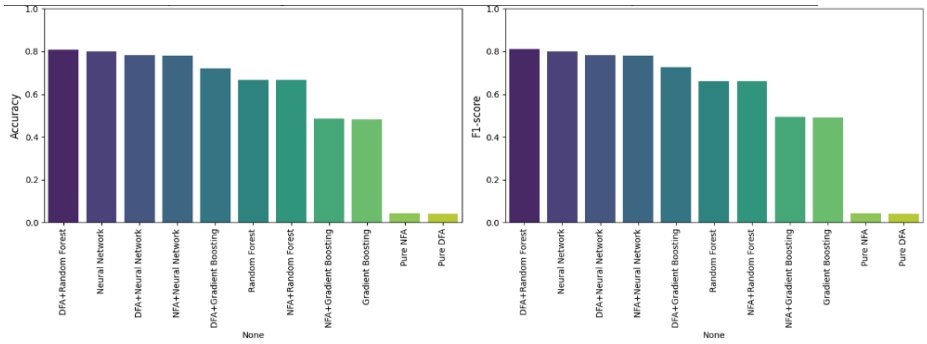


Figure 8 – Sign Recognition results

The table shows that pure finite automata models (Pure DFA and Pure NFA) demonstrate extremely low accuracy (about 4%), which indicates that they cannot be used independently for gesture recognition. Neural network models (DFA+NNN and NFA+NNN) showed slightly lower performance compared to pure neural networks. At the same time, models using purely neural networks show good performance with accuracy around 80% and F1-score around 80%. But the main success is shown by the DFA+RF model achieving the highest performance with an accuracy of about 80.9% and F1-score exceeding 81%. This indicates that the integration of structured features extracted using DFA with the power of ensemble methods can achieve significant improvement in the results.

Conclusion

The experimental results obtained confirm that integrating finite automata (FA) with machine learning (ML) methods improves the quality of sequence recognition compared to using pure methods. In the spam classification task, hybrid models such as DFA+NNN and NFA+NNN showed performance comparable to the best pure ML models, demonstrating that augmenting standard methods with binary features derived from finite automata allows for efficient extraction of specific patterns in text data. Similar findings were obtained in genetic sequence analysis experiments, where models using gradient bousting combined with KA showed the highest accuracy, and in a gesture recognition task, where hybrid models, especially those based on ensemble methods, outperformed the results of pure ML algorithms.

Comparing our results with data presented in a number of studies, the following can be noted. Early work in the application of finite automata to sequence recognition [4] showed high interpretability, but their accuracy was limited by the fixed structure of the models. Subsequently, studies such as Thomas et al. (2013) [5] and Wang et al. (2015) [6] have demonstrated that integrating ML methods with formal models can significantly improve recognition accuracy, compensating for the shortcomings of traditional finite automata. Our approach, which combines the capabilities of both deterministic and nondeterministic automata with state-of-the-art ML algorithms, demonstrates results comparable to or superior to those described in previous studies. Furthermore, studies by Aichernig et al. (2022) [7] and Kordestani et al. (2021) [3] emphasize the importance of adaptability and flexibility of models, which is supported by our experimental data, where hybrid models show stable improvements in accuracy and F1-score metrics. Meanwhile, the work of Ďuriš

et al. (2018) [1] notes scalability issues and complexity of finite automata configuration, which is also found in our studies, where the configuration of KA requires significant effort.

Finally, the study demonstrates that model transparency and interpretability remain important factors, and our approach integrating KA contributes to improved interpretability compared to pure ML models.

Based on this study, we can conclude that the integration of finite automata with machine learning algorithms can significantly improve the accuracy of sequence recognition, both in text classification tasks and in genetic data analysis and gesture recognition.

Despite the positive effects of integration, tuning finite automata requires additional effort and time, which can make the development of hybrid models more costly than pure ML methods, provided sufficient computational resources are available.

When choosing the optimal approach, the balance between accuracy improvement and the cost of developing and tuning hybrid models must be considered. In some cases, increasing the computational power for pure ML models may be more economically feasible.

For further research, we propose to develop methods for automated calibration of finite automata parameters, which will reduce time and cost when integrated with ML.

Exploring the application of the hybrid approach in other tasks, such as speech recognition, time series analysis, medical diagnostics, etc., where high quality of sequence recognition is critical.

Investigation of new methods for combining FA and ML to reduce complexity and increase scalability of the models, which will allow their application in big data environments.

Conducting additional comparative studies to evaluate the effectiveness of the hybrid approach in real industrial conditions, as well as analyzing the economic feasibility of implementing such systems.

References

1. Determinism and Nondeterminism in Finite Automata with Advice / P. Ďuriš et al // In Lecture notes in computer science. – 2018. – P. 3-16. https://doi.org/10.1007/978-3-319-98355-4_1.
2. Veanes M. Applications of symbolic finite automata / M. Veanes // In Lecture notes in computer science. – 2013. – P. 16-23. https://doi.org/10.1007/978-3-642-39274-0_3.
3. An introduction to learning automata and Optimization / J.K. Kordestani et al // In Intelligent systems reference library. – 2021. P. 1-50. https://doi.org/10.1007/978-3-030-76291-9_1.
4. Hong P. Gesture modeling and recognition using finite state machines / P. Hong, M. Turk, T.S. Huang // IEEE. – 2002. https://doi.org/10.1109/afgr.2000.840667.
5. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine / N. Thomas et al // Bioinformatics. – 2013. – № 29(5). – P. 542-550. https://doi.org/10.1093/bioinformatics/btt004.
6. Road network extraction: a neural-dynamic framework based on deep learning and a finite state machine / J. Wang et al // International Journal of Remote Sensing. – 2015. – № 36(12). – P. 3144-3169. https://doi.org/10.1080/01431161.2015.1054049.
7. Constrained training of recurrent neural networks for automata learning / B.K. Aichernig et al // In Lecture notes in computer science. – 2022. – P. 155-172. https://doi.org/10.1007/978-3-031-17108-6_10.
8. Privacy attacks to the 4G and 5G cellular paging protocols using side channel information / S.R. Hussain et al // in Proc. Netw. Distrib. Syst. Security Symp. (NDSS). – 2019. – P. 1-15.
9. Baray E. WLAN security protocols and WPA3 security approach measurement through aircrack-ng technique / E. Baray, N.K. Ojha // in Proc. 5th Int. Conf. Comput. Methodologies Commun. (ICCMC). – 2021. – P. 23-30.
10. Vanhoef M. Dragonblood: Analyzing the Dragonfly Handshake of WPA3 and EAP-pwd / M. Vanhoef, E. Ronen // 2020 IEEE Symposium on Security and Privacy (SP). – IEEE, 2020. – P. 517-533.

Acknowledgements. *The research data was sponsored by the Science Committee of the Minister of Science and Higher Education of the Republic of Kazakhstan (Grant No. of the research fund AP19678989 Intelligent video analytics and reporting on city streets surface and lighting).*

Ә.Қ. Қалпен [1], Э.Т. Матсон[2] , А.К. Жумадиллаева[1,3*], К.А. Дюсекеев[3]
[1]Astana IT University,
Қазақстан Республикасы, Астана қ., Маңғышлақ даң., 55/11 жыл, QazExpo С1 блогы
[2]Пердью Университеті,
Уэст Лафайетте, Индиана, АҚШ
[3]Л.Н. Гумилев атындағы Еуразия ұлттық университеті,
Қазақстан Республикасы, Астана, Қ. Сәтпаев көшесі, 2
e-mail: Zhumadillayeva_ak@enu.kz

АҚЫРЛЫ АВТОМАТТАР МЕН МАШИНАЛЫҚ ОҚЫТУДЫ ПАЙДАЛАНУ АРҚЫЛЫ ТІЗБЕКТЕРДІ ТАНУ

Тізбектерді тану көптеген пәндер бойынша маңызды міндет болып табылады. Ақырлы автоматтарға (КА) негізделген дәстүрлі әдістер деректерді құрылымдалған түрде ұсынуға және жоғары түсініктілікке қол жеткізуге мүмкіндік береді, алайда олардың икемділігі шектеулі. Қазіргі заманғы машиналық оқыту (ML) алгоритмдері жоғары дәлдікке қол жеткізеді, бірақ айтарлықтай есептеу ресурстарын қажет етеді. Осы парадигмаларды біріктіру күрделі тізбектерді тану тиімділігін арттыра алады. Бұл зерттеу тізбектерді талдау мәселелерін шешу үшін ақырлы автоматтар мен машиналық оқыту әдістерін біріктіруге арналған. Үш түрлі қолдану саласы қарастырылады: мәтіндерді жіктеу (спамды анықтау), Альцгеймер ауруымен байланысты генетикалық тізбектерді тану және сурет негізінде қимылдарды анықтау.

Әрбір сала үшін детерминирленген ақырлы автоматтар (DFA), недетерминирленген ақырлы автоматтар (NFA) және Random Forest, Gradient Boosting және көп қабатты перцептрондар (MLP) сияқты машиналық оқыту алгоритмдерін біріктіретін гибридті модельдер әзірленіп, тексерілді. Эксперименттік нәтижелер осы гибридті модельдердің дәстүрлі ML әдістерімен салыстырмалы өнімділікке қол жеткізетінін және кейбір жағдайларда дәлдіктің одан да жоғары болатынын көрсетті.

Спамды жіктеу міндетінде нейрондық желі модельдері ең үздік нәтижелер көрсетті, ал КА мен нейрондық желілердің комбинациясы ұқсас тиімділікті қамтамасыз етті.

Генетикалық тізбектерді талдау кезінде градиенттік бустингке негізделген модельдер ең жоғары дәлдікке жетті, ал ақырлы автоматтарды енгізу өнімділікті сақтап, түсіндіру қабілетін арттырды.

Қимылдарды тану міндетінде нейрондық желіге негізделген тәсілдер ең тиімді болып шықты, алайда ансамбльдік әдістермен ақырлы автоматтарды біріктіру дәстүрлі ML модельдерінен асып түсетін жоғары деңгейдегі болжамдық қабілеттілікке қол жеткізді.

Қорытындылай келе, ақырлы автоматтар мен машиналық оқытуды біріктіру тізбектерді талдау саласындағы перспективалы бағыт болып табылады. Болашақ зерттеулер модельдердің архитектурасын оңтайландыруға және күрделі құрылымдарды жоғары дәлдікпен тануды талап ететін басқа да салаларда қолдануға бағытталуы мүмкін.

Түйін сөздер: *ақырлы автомат, машиналық оқыту, тізбектерді тану, гибридті модельдер, генетикалық тізбектерді талдау, қимылдарды тану, мәтіндерді жіктеу.*

А.К. Калпен[1], Э.Т. Матсон[2] , А.К. Жумадиллаева[1,3*], К.А. Дюсекеев[3]
[1]Astana IT University,
Республика Казахстан, г. Астана, пр-т. Мангилик Ел 55/11, Блок С1 QazExpo
[2]Пердью Университет,
Уэст Лафайетте, Индиана, США
[3]Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан
Республика Казахстан, Астана, ул. К. Сатпаева, 2
*e-mail: Zhumadillayeva_ak@enu.kz

РАСПОЗНАВАНИЕ ПОСЛЕДОВАТЕЛЬНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ КОНЕЧНЫХ АВТОМАТОВ И МАШИННОГО ОБУЧЕНИЯ

Распознавание последовательностей является критически важной задачей во многих дисциплинах. Традиционные методы, основанные на конечных автоматах (КА), обеспечивают структурированное представление данных и высокую интерпретируемость, однако их гибкость ограничена. Современные алгоритмы машинного обучения (ML) демонстрируют высокую точность, но требуют значительных вычислительных ресурсов. Объединение этих парадигм может повысить эффективность распознавания сложных последовательностей. Данное исследование посвящено интеграции КА с методами ML для решения задач анализа последовательностей. Рассматриваются три различные области применения: классификация

текстов (определение спама), распознавание генетических последовательностей, связанных с болезнью Альцгеймера, и идентификация жестов на основе изображений.

Для каждой области были разработаны и протестированы гибридные модели, объединяющие детерминированные конечные автоматы (DFA), недетерминированные конечные автоматы (NFA) и алгоритмы машинного обучения, такие как Random Forest, Gradient Boosting и многослойные перцептроны (MLP). Экспериментальные результаты свидетельствуют о том, что данные гибридные модели достигают производительности, сопоставимой с традиционными методами ML, а в некоторых случаях обеспечивают более точные прогнозы.

При классификации спама нейронные сети показали наилучшие результаты, при этом комбинации КА с нейронными сетями продемонстрировали схожую эффективность.

В анализе генетических последовательностей модели на основе градиентного бустинга показали наивысшую точность, а интеграция КА позволила сохранить высокий уровень производительности при повышении интерпретируемости.

В задаче распознавания жестов наиболее эффективными оказались подходы, основанные на нейронных сетях, однако интеграция КА с ансамблевыми методами позволила добиться высоких прогностических показателей, превосходящих традиционные ML-модели.

В заключение, интеграция конечных автоматов и машинного обучения представляет собой перспективное направление в анализе последовательностей. Будущие исследования могут быть направлены на оптимизацию архитектур моделей и их применение в других областях, требующих высокой точности распознавания сложных структур.

Ключевые слова: *конечный автомат, машинное обучение, распознавание последовательностей, гибридные модели, анализ генетических последовательностей, распознавание жестов, классификация текстов.*

Information about the authors

Amirzhan Kuanyshuly Kalpen – master student, Astana IT University, Astana, Kazakhstan; e-mail: amirzhan103@gmail.com. ORCID: https://orcid.org/0009-0002-3545-4080.

Eric T. Matson – PhD, professor, Purdue University, West Lafayette, Indiana, USA; e-mail: ematson@purdue.edu. ORCID: https://orcid.org/0000-0001-9200-4903.

Ainur Zhumadillayeva* – candidate of technical sciences, associate professor, Astana IT University, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: Zhumadillayeva_ak@enu.kz. ORCID: https://orcid.org/0000-0003-1042-0415.

Kanagat Dyussekeyev – candidate of technical sciences, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan; e-mail: dyussekeyev_ka@enu.kz. ORCID: https://orcid.org/0000-0001-7691-2506.

Авторлар туралы мәліметтер

Әміржан Қуанышұлы Қалпен* – Астана IT университетінің магистранты, Астана, Қазақстан; e-mail: amirzhan103@gmail.com. ORCID: https://orcid.org/0009-0002-3545-4080.

Эрик Т. Матсон – PhD, профессор, Пердью Университет, Уэст Лафайетте, Индиана, АҚШ; e-mail: ematson@purdue.edu. ORCID: https://orcid.org/0000-0001-9200-4903.

Айнур Канадиловна Жумадиллаева* – техника ғылымдарының кандидаты, Astana IT University, Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан; e-mail: Zhumadillayeva_ak@enu.kz. ORCID: https://orcid.org/0000-0003-1042-0415.

Канагат Абетович Дюсекеев – техника ғылымдарының кандидаты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан; e-mail: dyussekeyev_ka@enu.kz. ORCID: https://orcid.org/0000-0001-7691-2506.

Сведения об авторах

Амиржан Куанышулы Калпен – магистрант Astana IT University, Астана, Казахстан; e-mail: amirzhan103@gmail.com. ORCID: https://orcid.org/0009-0002-3545-4080.

Эрик Т. Матсон – PhD, профессор, Пердью Университет, Уэст Лафайетте, Индиана, США; e-mail: ematson@purdue.edu. ORCID: https://orcid.org/0000-0001-9200-4903.

Айнур Канадиловна Жумадиллаева* – кандидат технических наук, Astana IT University, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан; e-mail: Zhumadillayeva_ak@enu.kz. ORCID: https://orcid.org/0000-0003-1042-0415.

Канагат Абетович Дюсекеев – кандидат технических наук, Евразийский национальный университет имени Л.Н. Гумилева, Астана, Казахстан; e-mail: dyussekeyev_ka@enu.kz. ORCID: https://orcid.org/0000-0001-7691-2506.

Received 20.02.2025
Revised 14.03.2025
Accepted 17.03.2025