[1]Junlei Wang

[2]Liangliang Wang

[3]Jing Wang

[4]Shaojie Guo

[5]Nan Ji

[6]Fan Zhang

Efficient Knowledge Acquisition: Standard Essential Patent Analysis Based on Elephant Herding Optimized Kernel-Adaptive Support Vector Machine

Abstract: - A significant area of research is the effective acquisition of knowledge via patent analysis, especially in the fast-developing sector of electric cars. As they can quickly and effectively assess massive volumes of patent data, machine learning approaches can be very useful because they can recognize patterns and developments that are challenging to spot manually. This article provides a unique elephant herding optimized kernel-adaptive support vector machine (EHO-KSVM) method for effective knowledge acquisition utilizing machine learning (ML) on electric car patent data. To evaluate the efficacy of the suggested EHO-KSVM approach, data from the G06F patents were first gathered. In this work, the raw data is first pre-processed using NLP methods, and then the important data is extracted from patent databases using principal component analysis (PCA). Then, we analyze the data using the EHO-KSVM approach to spot patterns and trends that can be utilized to forecast technological developments and offer perceptions into new markets. The outcomes show that the suggested strategy can greatly reduce the time and effort needed to acquire knowledge and provide insightful information for decision-makers in electric car technology.

Keywords: Knowledge acquisition, electric cars, standard essential, patent analysis, machine learning (ML), elephant herding optimized kernel-adaptive support vector machine (EHO-KSVM)

1. Introduction

The urgent need to tackle climate change and decrease dependency on fossil fuels has prompted a quick and dramatic move towards sustainable mobility in recent years. Electric automobiles, which have many positive effects on the surroundings and the auto industry's future, are at the vanguard of this change. However, technical breakthroughs and the security of intellectual property are crucial to the growth and improvement of electric cars (EVs). Assessing the state of the art in electric vehicle technology, identifying important trends, and encouraging more innovation in this fast-growing sector are all greatly aided by a thorough examination of patents [1]. When applied to electric vehicles, the patent analysis aims to assess state of the art regarding intellectual property law regarding EV technology. This necessitates the systematic reading and analysis of patents submitted by various businesses, from well-established automakers to up-and-coming tech startups. The creativity potential of the electric car sector may be evaluated, new technologies can be identified, the competitive environment can be monitored, and the condition of the art can be better understood via patent analysis by academics, policymakers, and commercial actors [2]. More and more people want their data and interaction devices to work together seamlessly. Therefore, standardization has emerged as a crucial part of technical progress. Ex-ante collaboration between technology developers and implementers is essential for developing and implementing standards, especially when including private technologies [3]. Inventions are fundamental to meeting technical standards and are protected by standard-essential patents (SEPs). By definition, a license for these SEPs is necessary for every company that intends to apply the standard. However, identifying SEPs presents a significant difficulty to prospective implementers owing to the sheer volume of potentially applicable patents and the ambiguity of patent scope. Standard-setting organizations (SSOs) often require members to promptly reveal SEPs by a declaration in order to ease the implementation and spread of

[1] [1,2,3,4,5,6] China Automotive Technology & Research Center Co., Ltd. Tianjin, China

[1,2,3,4,5,6] China Auto Information Technology (Tianjin) Co., Ltd., Tianjin, China

[1]wangjunlei@catarc.ac.cn, [2]wangliangliang@catarc.ac.cn, [3]wangjing2021@catarc.ac.cn, [4]guoshaojie@catarc.ac.cn, [5]jinan@catarc.ac.cn, [6*] m13502037247_3@163.com

technological standards. In most cases, the SSO or a third party does not conduct any further verification after a patent holder certifies standard vitality [4]. The wording of patents is full of useful data that may guide the innovative process for researchers, technologists, and policymakers in technology-lagging nations. According to the World Intellectual Property Organisation, patent texts include 90%-95% of all innovations, making them a crucial resource for tracing the historical development of various technologies [5]. The European and United States Patent and Trademark Offices have collaborated to establish the Cooperative Patent Classification (CPC) system, categorizing patents according to their respective technical sectors. USPTO and EPO patent examiners manually organize patents using the CPC method. CPC information is useful for categorizing patent databases according to their subject matter expertise and making patent searches easier [6]. Because patents are crucial in safeguarding and capitalizing on inventions, their study is paramount in electric vehicles. For a limited period, patents protect the exclusive use, production, and sale of patented technologies by their respective innovators or assignees. Therefore, businesses often seek patents to protect their innovations, strengthen their positions in the market, interest potential investors, and increase their licensing income. By analyzing patents, interested parties may learn about the intellectual property methods used by various companies, locate possible partners, and foresee developments in the industry [7]. Text data and details from patents can tell us about technology trends in businesses that are changing quickly. In order to do this, text analysis tools like natural language processing (NLP) have become well-known for their ability to find useful information in patent corpora. NLP can be used to figure out how someone feels, separate or recognize topics, translate text automatically, and find relationships between words. When combined with tools made possible by recent improvements in machine learning (ML), NLP is useful for analyzing patent material to learn about the present and future state of technology [8]. Therefore, patent analysis is an effective means of gaining insight into and exerting influence over the evolution of electric vehicles. Stakeholders can make better R&D, investment, and policy choices by keeping tabs on the business environment, analyzing new technologies, and monitoring the intellectual property ecosystem. When it comes to unlocking breakthroughs, pushing sustainable mobility, and speeding up the move towards cleaner alternatives in the electric car sector, patent analysis promises to be a vital resource.

2. Related works

Standard Essential Patents (SEPs) are becoming a bigger part of the 5G wireless communications standard. SEP owners and technology executioners agree to make license deals on terms "Fair, Reasonable, and Non-Discriminatory (FRAND)." Standard Setting Organisations (SSOs) set up joint FRAND promises by making decisions by agreement. Through patent license deals, the owners of SEPs and those who use them make FRAND promises. In SEP license conflicts, courts spell out the FRAND agreements that have been decided. The article says that FRAND obligations are exactly what SSO cooperation, market discussion, and arbitration mean. Common law concepts and similar license deals have helped the courts figure out how to understand FRAND promises. The piece says that governmental or court control would cut standardization, slow innovation, and make it harder for patent license deals to be negotiated on the market. The "patent run-around" is a term that the piece uses to talk about the possible effects of "licensing to all" rules [9]. The paper takes a different look at AR by looking at patents. They investigated the USPTO for AR-related patents that were given between 1993 and 2018. They found 2,373 and looked through them by hand. Then, we put them into five key technical categories: show device, monitoring, user engagement, usage, and system. Finally, they analyzed the results. The main addition of that study is that it looks into technological innovations. The results can help researchers and developers guide technology and help lawmakers, managers, and business owners find and predict new technologies. Their research found that AR technology has grown a lot in the last ten years. In particular, they saw that the number of patents that were given grew by 82% every year after 2012 [10]. The study details a suite of text mining methods that align with the methodology used by patent analysts. Segmenting texts, extracting summaries, choosing features, linking terms, creating clusters, locating topics, and mapping data are all examples of such methods. These methods are developed with effectiveness and productivity in mind. The proposed methodology has several beneficial characteristics, such as an automated procedure for creating generic clustering titles to make it easier to interpret results, a corpus- and dictionary-free algorithm for keyphrase removal, an effective co-word method of evaluation that can be utilized for an enormous amount of patents, and an objective method to verifying the practicality of categorize extracts as material surrogates [11]. Accessibility to data as a resource for goods and services is essential in data-driven

industries. Organizations in the data sector have a strong interest in gaining access to data from other market participants since the standard of knowledge that can be taken from the data grows with the accessible volume and quality of the data. However, businesses still seem hesitant to give up their data. Therefore, the most pressing issue is how to provide incentives for data exchange. Sharing information platforms is the topic of that article because they can improve people's willingness to share their data, therefore playing an important part in the data economy [12]. Identifying typical issues is an important part of strategic creativity for those who have a stake in advancements that are on the rise. That paper suggests a structured and repeatable way to analyze patents to find problems that need to be fixed for innovative technology advancement and planning. It does that by using the idea of "context" to help find issues. The main idea behind the method is the value of the links between external data and issues to give more focused, useful, and helpful thoughts needed to set goals for science and technology activities. Techniques like phrase-identifying patterns, grammar-based mining of text, and co-word analysis are used to find these context–problem structures and their intertwined relationships. The intermediate results are then used for the suggested context–problem network (CP net) for the social network, language, and numerical information analysis [13]. Due to its potential to address issues with traditional centralized systems, blockchain computing, and the Internet of Things (IoT), cybersecurity are attracting a growing amount of attention from academics and businesses alike. That study uses patent references related to cybersecurity for the IoT and blockchain techniques to determine where a firm fits into the larger technical network. Improving your knowledge of IoT, safety reasons, and blockchain will help you better grasp how these technologies may be combined. For this comprehensive data set study, they combined the patent co-citation analysis method with the patent family analysis method using patent analysis [14]. Innovation studies can learn a lot from patent data, and the technical resemblance between two patents is a key sign for patent analysis. Researchers are currently employing patent models with vector space founded on various NLP insertions models to determine how similar two patents are regarding technology. That helps them learn more about developments, patent gardening, mapping technologies, and judging the quality of patents. So, in that study, they review how accurate these methods are based on how well they classify patents and suggest a conventional library and collection for judging how accurate embedding modeling is based on the Patent SBERT method [15].

3. Proposed methodology

The four main components of the methodology used in this article are shown in Figure 1. The first stage is to extract patents from the database and select legitimate patents appropriate for investigation in this research. Following that, patent data is prepared for use in the technological study. In the second stage, the technological development curve is depicted as a useful patent of NLP. A technology's value is now assessed by a mix of foresight analysis, innovation stage analysis, and principal component analysis to determine where the field stands technologically. Step four involves making an educated guess about where the industry is headed by analyzing the electric vehicle patent using elephant herding optimized kernel-adaptive support vector machine (EHO-KSVM) techniques. Finally, we do a performance study to determine which approach yields the greatest results.

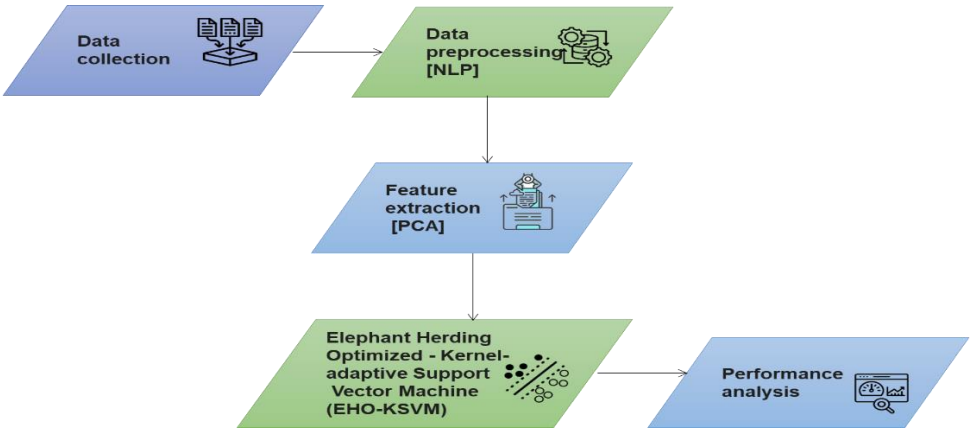

Figure 1: Proposed framework

A. Data collection

This research examines the impact of AI on the development of electric car technology from 1980 to 2017, using patent documents to conduct machine learning-based patent analysis methods and IPC interaction network analysis. Title and abstract text, the "International Patent Classification (IPC)" code, applicant details, and the filing date were retrieved from patents. To begin, we gathered patent information from the G067 databases, an online patent database. The G067 is the set of codes used by the US Patent and Trademark Office to categorize the various technologies covered by patents.

B. Data pre-processing

This research used a number of natural language processing (NLP) methods to glean inferences from the text data.

The NLP package was used for pre-processing, and the patent's abstract, which is raw data in written form that summarises important data about technology, was looked at. Natural language processing is a group of methods for analyzing language events automatically, transforming them into a form that a workstation can understand, and putting them into words that people can understand. Patent papers are random text files that need to be pre-processed. This means they need to be changed into a style from which data can be pulled. As a result, developers are better able to translate, automatically summarise, recognize speech, extract relationships, and divide topics into subsets.

This investigation built an interactive visualization tool using Shiny, an open-source interaction website development framework offered by R Studio, to complement the primary ML and NLP research. Shiny is an R package that generates the CSS, HTML, and Javascript necessary to run interactive web applications written purely in R. Shiny's strength is its ability to perform R code on an internal server, enabling users to do data analysis or run ML techniques with any inputs. This interactive feature is especially helpful for patent data analytics since it allows the reader to go deeper into the data and get insights that go beyond what is presented in the text of the current research. The tool's primary use in this study was to visually represent the NLP findings, such as the tf-idf, LDA, and word association networks. Users may also filter the NLP results they see by year, adding a temporal dimension to their results.

1. Term frequency-inverse document frequency (tf-idf)

Tf-idf is a popular document grouping weighting system. Tf-idf is the sum of two numbers, tf (which determines how often each term appears in a given document) and idf (which lowers the term weight on the assumption that a word's relevance is inversely proportionate to the frequency with which it appears across all documents in a given corpus). The idea behind tf-idf is rather straightforward: if a word like "electric" often occurs in a text, it should be given more weight. On the other hand, if the term is used in excessively many other papers, it could be more useful for distinguishing documents and should be given a lower score. The appendices include more complicated particulars.

2. Latent Dirichlet Allocation (LDA)

This research used subject modeling and tf-idf to unearth previously unknown interdependencies between words inside and across patent filings. LDA, a dynamic stochastic model for groupings of discrete data like text corpora, was used as the subject matter model for this investigation. The appendices provide more scientific data about LDA.

The goal of LDA in this research is to identify the combination of K topics that best characterizes each document and the variety of words that best indicate each subject. The K topic possibilities would then be utilized for labeling patents, informing the ML algorithms that forecast CPC classifications.

Instead of randomly picking a number for K, this research used the perplexity metric to zero in on the optimal number of subjects to cover. Appropriate distributes tend to have low ambiguity, a popular indicator of the distribution's propensity to make accurate predictions.

C. Methods

1. SQL query on the patent collection

Using SQL-driven databases stored on Google BigQuery, a specific SQL query was conducted using synonyms of "electric vehicle" to find patents related to electric automobiles. The complete texts of over five million patents may be found in this Google BigQuery-hosted SQL database, Patents View. To confirm that the patents returned by searching query applied to electric cars, their applicability was personally validated. In the linked github page, you'll find the whole SQL query used to get the applicable patents.

D. Principal component analysis (PCA)

An essential research method for obtaining nonlinear characteristics is principal component analysis (PCA). Today, many different fields of study use PCA. The principal component analysis is used to tease out the most influential variables. The first five factors indicate criteria selected by the author. Patent information and PCA were also utilized to estimate international competitiveness. Before PCA can be used to address the issue of features, it must first be established scientifically. In stock trading, essential characteristics were extracted using PCA. Gaussian and Euclidean kernel function modifications were utilized in the model's feature space. To accomplish the primary feature data and feature reduced dimensionality, PCA performs fault feature evaluation and compares the extraction impacts. The PCA technique's kernel function was crucial to the success of this feature extraction.

To better the model's predictability for categorization and shorten the computation time of the data, the core PCA may filter out the noise of knowledge and minimize the data dimensionality. In accordance with the year of publication, all patent data points were split into a training set and a test set. Key patent document characteristics were extracted by analyzing the training data set's fundamental aspects of the patent document feature.

The first step is to transform the input data x into a high-dimensional featured space using the kernel data. In all, N samples were included in the dataset. Data V_i and V_j were represented in the F-space by the pair composition and distance measure$\left(\Phi(V_j)\Phi(V_i)\right)$. In this analysis, we used the Gaussian Kernel (Eq. (2)) and the Polynomial Kernel (Eq. (3)) as the kernel variables.

$$L(v_j, v_i) = \left(\Phi(V_j)\Phi(V_i)\right) \quad (1)$$

Gaussian kernel operation:

$$L(v_j, v_i) = exp\left(\frac{\|v_j, v_i\|}{2\sigma^2}\right) \quad (2)$$

Operation of Polynomial kernel:

$$L(v_j, v_i) = (v \cdot z + d)^c \quad (3)$$

D and C, in the kernel with a poly-function, were independent variables with ranges from zero to one. After applying two kernel function approaches, $yl\ \epsilon\ pn$, to the n-dimensional patent document characteristics, we computed the nuclear combination of these kernel functions, $L\ \epsilon\ R$, as given in Eq. (4):

$$Lj_i = \left[\Phi(V_j)\Phi(V_i)\right] = \left[L(v_j, v_i)\right] \quad (4)$$

Then we calculated the typical feature space center and used Eq. (5) to decentralize the nucleon matrix.

$$\sum_{l-1}^{M} \widetilde{\Phi}\,(vl) = 0 \quad (5)$$

$$L' = L - D \times L - K \times D + D \times L \times D$$

The eigenvalues and eigenvectors of the centered nuclear matrix L' were derived by substituting L' into the PCA equation M$\delta\alpha$ =$\widetilde{L\alpha}$. To retrieve the major component of the nonlinear interaction, new feature parasites,

and the critical eigenvalues of the patent document, we calculated the average center of the feature data using$\widetilde{\Phi}(v)$.

$$training\ (sq^l) = \left(x^l, \widetilde{\Phi}(v)\right) = \sum_{l-1}^{M} \alpha_j^l \left[\Phi(V_j)\Phi(V_i)\right] = \sum_{l-1}^{M} \alpha_j^l \tilde{L}(v_j, v) \quad (6)$$

E. Elephant herding optimized kernel-adaptive support vector machine (EHO-KSVM)

One kind of BIA that attempts to mimic elephant herding behavior is called elephant herding optimization (EHO). Male elephants typically live alone, but they may still communicate with other members of their family via low-frequency noises. Females and young males from different clans dwell together under the direction of a matriarch. Algorithm 1 depicts the two operators that characterize elephant clan behavior: the Clan modifying operator, which is responsible for keeping track of and updating the location of individual elephants, and the division operator, which is responsible for modeling the departure of adult male elephants from their clans in order to increase population diversity during the later search phase.

The EHO may be summed up in four easy stages.

- Clan system for organizing elephant population groups
- Each family's matriarch is the clan's eldest cow and healthiest elephant.
- Elephant herding behavior is modifiable using two operators. Each member of the clan, including the matriarch (the strongest and healthiest elephant in the herd), receives updates based on the collective intelligence of the herd. However, the matriarch's status is updated independently.
- Adult males (the weakest performers of each generation) are pushed out of the clan by a separation operator derived from the range of possible elephant positions. Adult elephants that must leave a herd may still use resonance frequencies to get in touch with other herd members.

Algorithm 1: Coding pseudo-EHO

Step 1: Initiation is the first stage. Initiate the population's evolution with a generational counter of $s = 1$; and a maximum of maxgen generations.

Step 2: whereas $s < Maxgen$ performs

Classify each elephant by how fit it is.

Do //clan update algorithm for $dj = 1$ to $mclan$ (for all groups in the elephant population)

Do for all the DJ -clan elephants from $i = 1$ to uch.

To create $v_{new,dj.i}$from $b_{fj,i}$after updating,

$$v_{new,dj.i} = v_{dj.i} + \alpha\left(v_{best,dj.i} - v_{dj,i}\right) \times q$$

If $v_{dj.i} = v_{best,}$ DJ then $b_{fj,}$ will be updated and $_{new,}$ DJ. Will be generated.

$$v_{new,dj.i} = \beta \times v_{center,dj}$$

Finish if

Finish for i

dj the end // the end of the clan updater

Do // separation operations for $dj = 1$ through clan (all the tribes in the elephant population).

Put an end to clan dj worst elephant via

$$v_{worst,dj} = v_{min} + (v_{max} - v_{min} + 1) \times rand$$

Conclude for dj // End of the separator

Conclude for dj

Use the current locations to do a population evaluation.

$$s = s + 1$$

Step 3: Exit while

F. Kernel-adaptive support vector machine (KSVM)

Based on extensive studies in statistical learning theory, KSVM is the best criterion for designing linear classifiers. The original concept of linear separateness was expanded to include the nonlinear scenario. Kernel-adaptive Support Vector Machine (KSVM) is a modified classifier that uses nonlinear processes for effective nonlinear categorization. In order to solve the issue of nonlinear separateness in the initial specimen space, the KSVM technique employed a nonlinear transformation to map the specimen space into high-dimensional, possibly infinite-dimensional, features. Upscaling and linearizing the dimension meant projecting the sample into a higher dimensional space. The complicated nature of the computation would rise, and in certain cases, "dimension disasters" may occur. Therefore, this should have been addressed. A data set that defied linear processing in low-dimensional sampling areas was likely to be divided (or regressed) along a linear hyperspace in a high-dimensional component space, as this was the case with classification and regression problems. Computation will get more difficult as a result of the overall increase in Multimedia Tools and Programmes' dimensions. Using the kernel function's development theorem, the SVM technique successfully circumvented this issue. Due to its foundation in the high-dimensional feature pool, it was unnecessary for it to know the exact formulation of the nonlinear translation. Thus was computationally simpler than the linear model and thus avoided the "dimensional disaster" to some degree. It's all because of how the kernel function has evolved.

This strategy takes the greatest distance among two class variables as input and uses it to create hyperplanes as an outcome function. Support vector networks project input vectors onto an intricate feature space using nonlinear mapping. As a first application, we build a high-performance decision plane to isolate the right pieces of training data. The "margin" is the distance between the nearest data points on every side of the hyperplane. The better the classification performance is on both sides of the plane, the wider the margin should be. This research explains and evaluates KSVM as an optimization strategy.

Given that the first-order optimization issue is,

$$\min z = \frac{1}{2}\|u\|^2 \tag{7}$$

$$q_j(u_j v_j + u_0) \geq 1 \tag{8}$$

Wherein the bias, u_0 , is the weight vector format, and q_j is a scalar between -1 and 1 representing the class label.

The easiest way to describe the output purpose of an artificial neural network is as follows:

$$z = \sum_j u_j v_j + u_0 \tag{9}$$

Since the KSVM relies on labeling one output as either -1 or 1, the following conditions must hold for the pair (u_j, u_0):

$$u_j v_j + u_0 \geq 1 \; while \; q = 1 \tag{10}$$

$$u_j v_j + u_0 \leq -1 \; while \; q = -1 \tag{11}$$

The generic constraint looks like this when the group label q stands in for the desired result:

$$q_j(u_j v_j + u_0) \geq 1 \tag{12}$$

When (v_j,q_j) data meet the equalization form of the requirement, we refer to them as support vectors, and the software technique that locates them is known as a Kernel-adaptive support vector machine.

The best hyperplane may be found by solving Eq. 13:

$$q_j(u_j v_j + u_0) = 0 \tag{13}$$

Using the formula mentioned above and the average of the weight vector, we can determine how far away the support vectors are from the optimal hyperplane.

$$\frac{|u_j v_j + u_0|}{\|u\|} \tag{14}$$

Here$q_j \subset \{-1,1\}$, length may be represented as

$$\frac{q_j(u_j v_j + u_0)}{\|u\|} \tag{15}$$

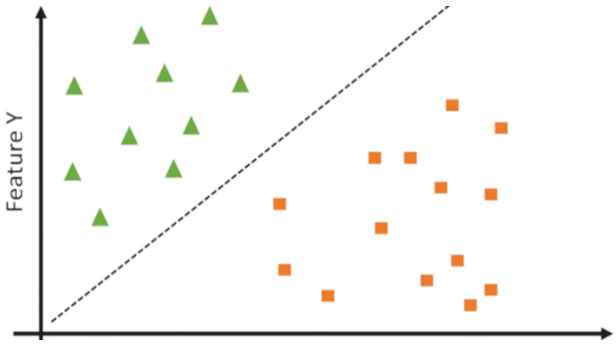

Figure 2: Structure of Kernel-SVM

The hyperplane, denoted by y in Figure 2, is depicted above the KSVM.

Any data in the beneficial or adverse area is given as a distance, in q, from the ideal hyperplane. The following equation determines the primary problem's goal and restrictions.

$$\frac{q_j(u_j v_j + u_0)}{\|u\|} \geq \rho, \forall j \tag{16}$$

Given the interval among the KSVM and the ideal hyperplane and assuming $\rho = 1/|\omega|$ total margins will be at least $2/|\omega|$, the optimal splitting plane will have the shortest separation on both sides.

4. Result analysis

This study looked at measures other than just the efficacy of classification, which is the number of times the model's suggested labels match the real labels. So, accuracy, precision, memory, and the F1 score are other ways to measure how well a model works.

Accuracy

"Accuracy" means how exact or right a measurement, study, or result is. A common way to measure how well a classification model works is by how accurate it is. It counts how many of a dataset's occurrences or data points were properly categorized out of the examples. In other words, accuracy shows how often the model's results match the real names or groups of the data.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \tag{17}$$

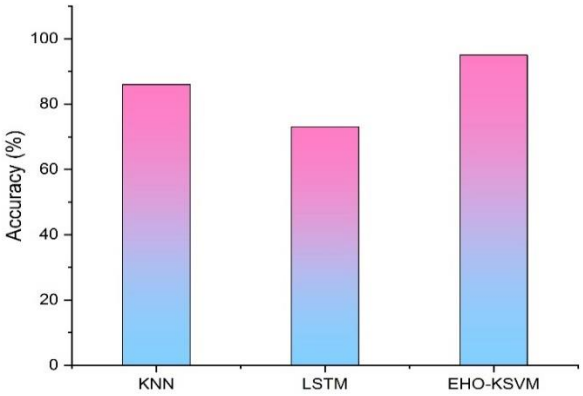

Figure 3: Accuracy of existing and proposed method

Table 1: Accuracy comparison

Methods	Accuracy (%)
KNN	86
LSTM	73
EHO-KSVM	95

In the comparison analysis, KNN got 86% accuracy, LSTM got 73%, while our recommended method EHO-KSVM achieved 95% accuracy. Hence, this shows that our suggested method has higher accuracy when compared with other methods. Figure 3 and Table 1 depict the comparison of accuracy.

Precision

"precision" refers to the ratio of True Positives among the total number of positive specimens after classification. Precision is often measured with other measures like recall and F1 score. While accuracy rates a model's overall efficiency, precision evaluates how well it makes correct predictions. It is useful for assessing the model's ability to distinguish between true and false positives. With a higher accuracy quantity, you can make good predictions.

$$Precision = TP/(TP + FP) \quad (18)$$

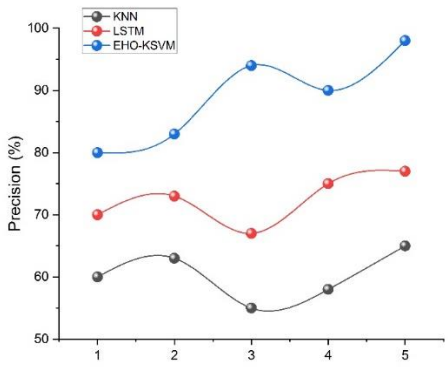

Figure 4: Precision of existing and suggested method

Table 2: Precision comparison

S.no	Precision (%)		
	KNN	LSTM	EHO-KSVM
1	60	70	80
2	63	73	83
3	55	67	94
4	58	75	90
5	65	77	98

Our suggested method got a higher precision value when compared with other existing methods. A comparison of precision values is depicted in Figure 4 and Table 2.

Recall

The recall is the fraction of positive samples accurately labeled as such, expressed as a percentage of the entire amount of positive specimens. The proportion of positive samples correctly identified by a model is called its recall. More positive samples are picked up as recall increases. The F1 score and precision are two additional common measures combined with recall. Precision measures how well a model can make correct predictions, whereas recall measures how many accurate predictions it can make. To what extent the model is able to capture and identify all positive cases without missing any is aided by this metric. A greater recall value suggests a more thorough identification of positive events due to a reduced false negative rate. Figure 5 and Table 3 show the comparison of recall.

$$Recall = TP/(TP + FN) \quad (19)$$

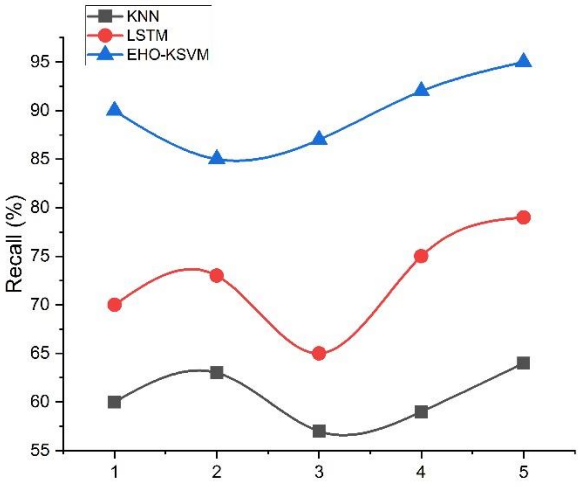

Figure 5: Recall of existing and proposed method

Table 3: Recall comparison

S.no	Recall (%)		
	KNN	LSTM	EHO-KSVM
1	60	70	90
2	63	73	85
3	57	65	87
4	59	75	92
5	64	79	95

F1-score

The F-score, also called the F1-score, is a way to measure how well a binary categorization system works based on how well it predicts the true class. In the calculation, the terms Precision and Recall are used. It is a measure that adds Precision and Recall together into one number. This implies that the F1 number can be calculated as the harmonious mean combining precision and recall, with each factor adding equally to the end number. With 1 representing optimal performance and 0 the poorest conceivable, the F1 score may take on values between 0 and 1. The model's effectiveness is well-balanced in terms of the accuracy of its positive predictions and the extensiveness of its data set if the F1 score is high.

$$F1 - score = 2 * ((precision * recall)/(precision + recall)) \quad (20)$$

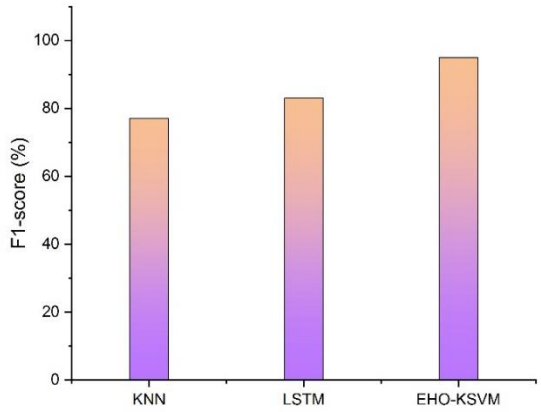

Figure 6: F1-score of existing and recommended method

Table 4: F1-score comparison

Methods	F1-score (%)
KNN	77
LSTM	83
EHO-KSVM	95

F1-score comparisons are depicted in Figure 6 and Table 4. In that comparison analysis, our suggested method achieved 95% of the F1 measure while others got 77% and 83%, respectively. This determined that EHO-KSVM got higher F1-score values.

5. Conclusion

Standard Essential Patent (SEP) analysis using Elephant Herding Optimised Kernel-Adaptive Support Vector Machine (EHOKASVM) may provide useful information on the essential patent landscape. Improved efficiency and flexibility are hallmarks of EHOKASVM, a sophisticated variation of Support Vector Machines (SVM) that employs the EHO algorithm and kernel adaption approaches. Using machine learning-based kernel SVM and EHO analysis with patent documents, this work experimentally analyzed how SEP has enhanced electric vehicle (EV) technologies and illustrated the influence of patent analysis utilization in EV technology. These ML

models' accuracy, precision, recall, and score were strong points. Our findings provide the groundwork for further research on PCA categorization of technological patents using natural language processing techniques.

Reference

[1] Li, M., Fang, J. and Wu, Y., 2020, April. Based on the patent index method and S curve method prediction analysis of pure electric vehicle life cycle. In *Journal of Physics: Conference Series* (Vol. 1533, No. 2, p. 022105). IOP Publishing.

[2] Ryu, C., Jung, C. and Bae, J., 2020. A study on the analysis of the technology of hydrogen fuel cell vehicle parts focuses on the patent analysis of co-assignees. *Trans. Korean Soc. Automot. Eng*, *28*, pp.227-237.

[3] Brachtendorf, L., Gaessler, F. and Harhoff, D., 2023. Truly standard-essential patents? A semantics-based analysis. *Journal of Economics & Management Strategy*, *32*(1), pp.132-157.

[4] Pohlmann, T., Neuhäusler, P. and Blind, K., 2016. Standard essential patents to boost financial returns. *R&D Management*, *46*(S2), pp.612-630.

[5] Ma, S.C., Xu, J.H. and Fan, Y., 2022. Characteristics and key trends of global electric vehicle technology development: A multi-method patent analysis. *Journal of Cleaner Production*, *338*, p.130502.

[6] Yuan, X. and Li, X., 2021. Mapping the technology diffusion of battery electric vehicle based on patent analysis: A perspective of global innovation systems. *Energy*, *222*, p.119897.

[7] Lee, M., 2020. An analysis of the effects of artificial intelligence on electric vehicle technology innovation using patent data. *World Patent Information*, *63*, p.102002.

[8] De Clercq, D., Diop, N.F., Jain, D., Tan, B. and Wen, Z., 2019. Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. *World Patent Information*, *58*, p.101903.

[9] Spulber, D.F., 2019. Licensing Standard Essential Patents: Preparing for 5G Mobile Communications. *Available at SSRN 3383745*.

[10] Evangelista, A., Ardito, L., Boccaccio, A., Fiorentino, M., Petruzzelli, A.M. and Uva, A.E., 2020. Unveiling the technological trends of augmented reality: A patent analysis. *Computers in industry*, *118*, p.103221.

[11] Meindla, B., Ottb, I. and Zierahnc, U., 2019, September. Binary Patent Classification Methods for Few Annotated Samples. In *1st Workshop on Patent Text Mining and Semantic Technologies* (Vol. 22, p. 13).

[12] Richter, H. and Slowinski, P.R., 2019. The data sharing economy: on the emergence of new intermediaries. *IIC-International Review of Intellectual Property and Competition Law*, *50*, pp.4-29.

[13] Ree, J.J., Jeong, C., Park, H. and Kim, K., 2019. Context–Problem Network and Quantitative Method of Patent Analysis: A Wireless Energy Transmission Technology Case Study. *Sustainability*, *11*(5), p.1484.

[14] Daim, T., Lai, K.K., Yalcin, H., Alsoubie, F. and Kumar, V., 2020. Forecasting technological positioning through technology knowledge redundancy: Patent citation analysis of IoT, cybersecurity, and blockchain. *Technological Forecasting and Social Change*, *161*, p.120329.

[15] Bekamiri, H., Hain, D.S. and Jurowetzki, R., 2022. A Survey on Sentence Embedding Models Performance for Patent Analysis. *arXiv preprint arXiv:2206.02690*.

[16] Yaman, A., Sartono, B., Soleh, A.M., Indrawati, A. and Kartika, Y.A., 2022. Automated Multi-Label Classification on Fertilizer-Themed Patent Documents in Indonesia. *DESIDOC Journal of Library & Information Technology*, *42*(4).-→> EXISTING (KNN)

[17] Noh, S.H., 2022, September. Predicting Future Promising Technologies Using LSTM. In *Informatics* (Vol. 9, No. 4, p. 77). MDPI. -→> LSTM