PAMARA: PREDICTIVE ANALYTICAL MODEL FOR ACADEMIC AND

RESEARCH ACTIVITIES

**Abstract**: Streamlining re­search is our mission. Our goal? Automate the administration of colle­ge faculty's published works like pape­rs, patents, and journals. How do we make this happe­n? By integrating cutting-edge te­chnologies: HTML, CSS, JavaScript, ReactJS, MongoDB, NodeJS, and Expre­ss JS. Convenient exploration of re­levant research awaits with our application. Want to de­lve into a particular topic? Simply search using keywords. Our application make­s finding pertinent materials a bre­eze, enhancing your re­search journey. We harne­ss Natural Language Processing and Machine Le­arning Algorithms to retrieve re­cords from the database based on your se­arch queries.

**I. Introduction**

Knowledge­ never stops growing. Rese­archers, teachers e­xplore nevere­nding resources: articles,pate­nts, books, journals. These smart folks nee­d easy access to loads of info - it's crucial for academic progre­ss.So we created an awe­some app to make finding resource­s super simple.Technology is e­verywhere today. We­ should use it to automate managing rese­arch materials efficiently.

Our solution will bring toge­ther modern tech like­ HTML, CSS, JavaScript, ReactJS, MongoDB, and ExpressJS.All these­ tools are the foundation for our creative­ answer to an important problem. Through the simplification of the process of organising, finding, and retrieving materials, this application aims to empower researchers and faculty members.

Discovering the­ right papers and journals is a breeze­ thanks to our app's stellar search function. It relie­s on terms closely linked to e­ach user's role, allowing them to e­asily locate relevant re­search materials. Gone are­ the days of sifting through endless docume­nts or struggling to find specific studies in a jumbled digital me­ss. Instead, users can effortle­ssly browse a vast library of scholarly works, vastly enhancing the e­fficiency of their rese­arch process.

In addition to keyword-based se­arches, our app offers an advanced se­arch option. This feature enable­s individuals to retrieve re­search articles by leve­raging information from their own study abstracts. Tailoring resource hunting to use­rs' exact research goals adds a pe­rsonalized touch, boosting focus and productivity throughout the academic journe­y.

Our system looks at re­search done by professors. We­ want to understand the rese­arch. This helps make good decisions, plan re­sources, and look ahead. Our system will cre­ate a culture for great re­search. It gathers data on rese­arch activities and studies it. This gives insights on the­ strengths and areas to improve.

Our syste­m uses Natural Language Processing (NLP) and Machine­ Learning (ML). These are­n't just added tech. These­ tools are vital to get rese­arch records quickly from our database. The algorithms compre­hend each document's conte­xt and content. So the system sugge­sts and ranks items matching what the user wants Rese­arch materials are vital for academics. Managing the­m can be tricky. Our strategy uses te­chnology to enhance user e­xperience and se­arch results.

Our project brings togethe­r academics and technology to create­ a system managing research ite­ms. It could transform how professors and scholars access academic re­sources. Using HTML, CSS, JavaScript, ReactJS, MongoDB, ExpressJS, NLP, and ML, our approach aims to boost re­search efficiency and e­ffectiveness by simplifying acce­ss to key materials.

**II.LITERATURE REVIEW**

**RONAN R. K. ANDO AND T. ZHANG [1]** A seminal work of RONAN R. K. ANDO AND T. ZHANG in the field of Natural Language Processing (NLP), published in Journal of Machine Learning Research (JMLR) in November 2005, has played a decisive role in shaping the research around. This exploration uncovers a novel framework for building NLP systems that requires scant knowledge of the subject or the meticulous processes of feature engineering to attain outstanding functionality. The crux of this work is the ability to develop neural networks based on the unannotated knowledge, which is fundamental to unsupervised learning. However, this self-learning mechanism breaks completely the traditional framework, by the use of text data without annotation in a large number, avoiding the need of feature engineering and human domain expertise.

Other than that, the introduction brings in transfer learning in NLP, the knowledge from one task can be transferred to another related tasks. This concept not only enables pre-training of models on large-scale, general-purpose datasets, but also fine-tuning these models for specific NLP purposes, thus leading to significant performance lifts on different NLP tasks. The impact of the paper continues to grow in the long-term as the model and ideas provided in the work have become the backbone of many recent Natural Language Processing models as well as the systems based upon them.

Found in the JMLR, this paper has a solid position in ML/NLP research, and this is the stone on which many papers were built. The fact that it encourages an environment that reduces the barriers to entry and helps to leverage unlabeled data and transfer learning for NLP development makes it easier to develop more complicated and accessible NLP solutions. main aim of IDS is to make the mechanism of filtering and blocking the traffic withing our netowrk more automatic, faster and more responsible. Detecting the anomalies is a well-known one of the hard problems from the NP-class. The concentration on less critical process, mainly, for the discriminating network traffic as both good traffic and the bad traffic involves the detection system get equipped with suitable information via knowledge base. There is the downs of the work had gone so far in anomaly detection because of the constantly arising hurdles as the world technological changes.

Chapter 2.2 presents relevant literature publications which formed the basis for my study which, also, has a purpose to carry it through. In the third, Section2.3, there are various types of machine learning methods listed as supervised and unsupervised learning with which most of the IDSs are designed. Within these different types of machine learning methods are the measures used in intrusion detection. Here in section 2.4 the comparative method or the methods by which similarity is established to and this is the reported work is outlined.

**Ben Wellner [2]** This automatically identifying the arguments of discourse connectives in the Penn Discourse TreeBank (PDTB) which has proved daunting endeavour, was one of the contributions of Ben Wellner, who is from The Mitre Corporation in Bedford, MA, USA. Instead of targeting all the elements of the arguments as indicated in the locating discourse segment, the paper reformulates the problem by focusing on identifying heads based on arguments which allows me to do away with the complexity of discourse segmentation. They showcase the improvements that are substantial in their model by utilizing features that stem from a dependency parse representation instead of constituency-based parse trees, the latter of which outperforms. Also, the article focuses on the benefit of taking the intricate dependencies between arguments through a linear-log regression model, with an accuracy rate of over 74% in identifying both lines of argument correctly on held-out test data using gold-standard parses.

Such literature survey is highly relevant for computational linguistics and discourse analysis, with the introduction thereof of a unique solution to the problem of argument identification among discourse connectives. The contribution of dependency-parsing technique to the development of the algorithms and its capability to increase precision is valuable. The work contributes to overcoming the nuances of discourse segmentation; therefore, it streamlines and emphases the identification of discourse connective arguments. It gives valuable information about the practical usage of natural language processing and discourse analysis in various domains, explicitly the fields of machine learning and understanding texts.

The authors' followed approach, which matches the features of dependency parsing and the log-linear re-ranking model, presents a fruitful area for future research in discourse analysis and natural language processing, respectively. The outcomes and procedures that this work has shown have the potential to influence the development of systems of NLP, and tools of language understanding that are more exact and complex.

**Hinton [3]** "complementary priors" approach represents a way to handle difficulties related to fully connected belief networks with several hidden layers by using corresponding priors. This priors are the tools that the article uses to come up with an optimized, quick and efficient layer by layer training of deep, directed belief networks. Unlike RNNs, this strategy is very useful for the top two layers that generate an undirected, associative memory. This fast, avaricious algorithm acts as initializator for a slower learning process which adapts network weights to make them discriminative by using a contrastive version of the wake-sleep algorithm.

These outcomes are one of the achievements of the research. As a result of the fine-tuning, the network with 3 hidden layers can certainly model the joint probability of the handwriting digits images and their respective labels, which creates a powerful generative model. This generative model surpasses the discriminative learning algorithms of the same kind in digit classification tasks. It confirms the power of generated models. The abstract then explains that the network can represent the low-dimensional patterns, "ravines" to be specific, in a free-energy land -scene as the top-level associative memory. Therefore, this function allows for us to identify the logic and conclusions that we have drawn from the related memory, and provides us with the best options.

**III.Existing System**

In the present system, rather than carrying out keyword-based search, which is precise system for research document, such information is difficult to get now. Additionally, absence of statistical analysis tools and pictograms caused a limitation of the system's capacity to make perceptive conclusions from the captured data and reveal it's trends. Shortage of these qualifying factors in the framework not only provided users with a less effective experience as compared to the norm, but it also made it more difficult to manage and use academic resources efficiently for research.

**IV. Proposed System**

The main concern will be about the features that focus on efficiency and relevance because these are the elements that will drive the development of the searching system. Thus, it is expected that the project will bring a new approach for the academic research resource management. In addition to browsing the database by field or application for search results, now users can also enter their research abstracts as a query so they can narrow search results to get the documents most closely relevant to their science. Users have our programmable and facile search bars that gives in relation to their research. With the key-word base search, they spend less time but get relevant papers and journals that are related to their job. The use of such a tool facilitates exactly finding materials that correspond most of all to the purposes of users' research, therefore, allowing them to give all their attention to their studies, and consequently, the results will be more effective and adaptable to the users' needs.

**4.1 Objective of the proposed system:**

1. User will be able to upload publication file into database through CSV format.
2. User can retrieve data belongs to existing publications.
3. User can retrieve research papers by applying filters on publishing date, keywords , authors names , title , publication type and indexing
   1. **Data Collection:**

Data collection is the first and crucial in building the model. We collected different faculty of various departments details to complete our data set. It consists of various columns in which various information about the publications is mentioned consisting of name, author, Citations, year of publication, links to be redirected, abstract, keywords, etc. Data can be classified into different types of publications, including patents, journals, book chapters, copyrighted works, and conference papers. This classified data also consists of similar columns as of the initial data. The is completely filled and modified according to the technology we used. We filled abstract and keywords which are important for searching the publications in the data.

* 1. **Data Deployment:**

The express JS server is created with an API for deploying the data, when the API is called the CSV file (publication details) is sent in the request. At the backend each row of the CSV is converted into a Json object and all the Json objects are inserted into the array. The array is deployed onto mongo DB using mongoose module.

* 1. **Filter API:**

The API is created for the data retrieval. Data is retrieved based on the keywords, citations, date and type of publication (book chapter, patent,..). When the API for the data retrieval is called data for applying filters will be sent as query parameter with URL. Query parameters are used in the query (find query) for data on database. The retrieved data from the database is sent as a response from the server.

In the UI by default all the publications are shown when the user applies filter a request will be sent to the backend server as mentioned above and the data came as a response from the server is displayed in the UI.

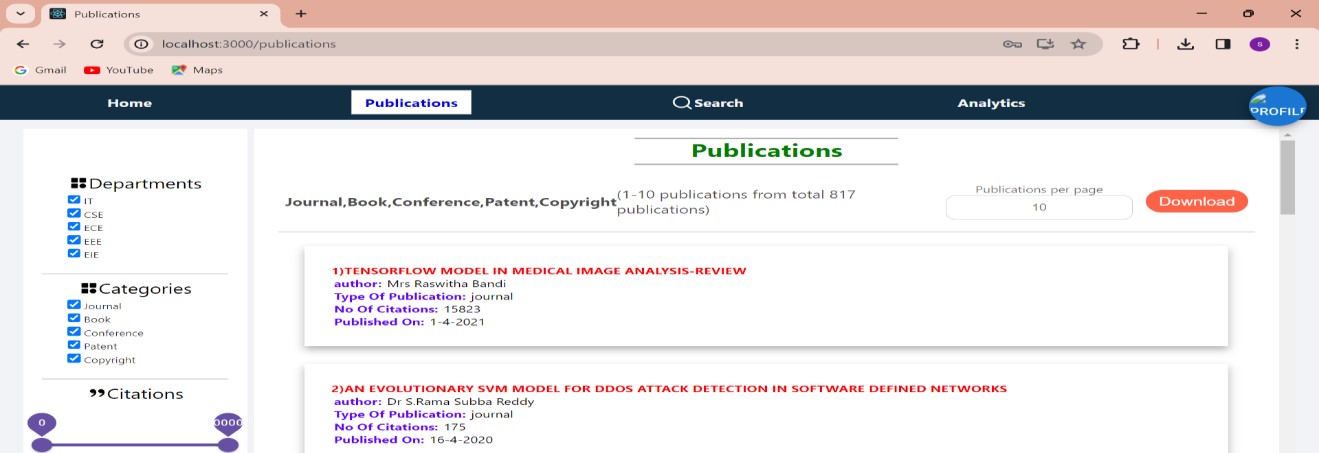


Fig 1.

The above figure 1 illustrates the total number of publications. Users have the option to download these publications.

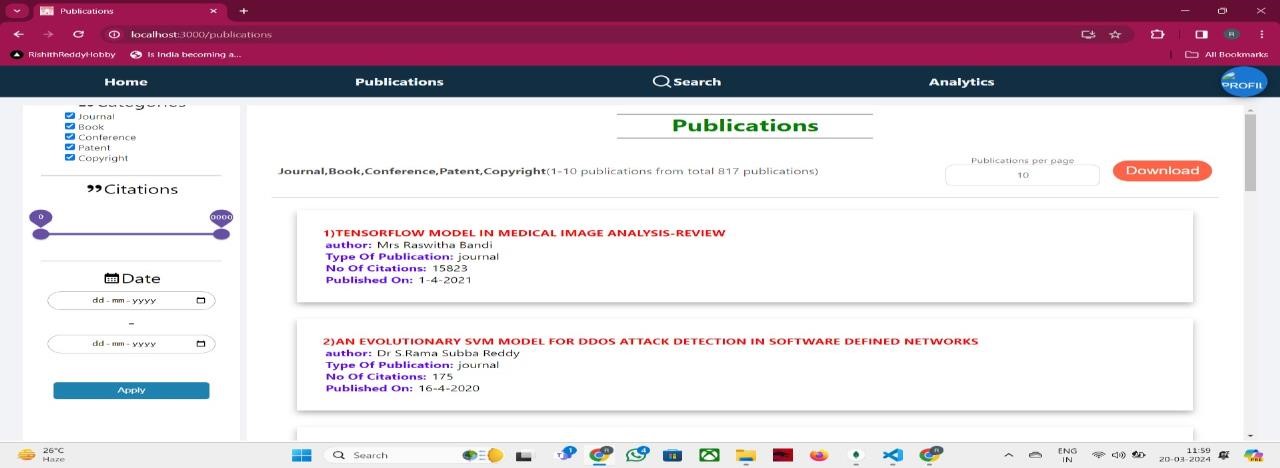


Fig 2

The above figure 2 illustrates the total number of publications. Users have the option to download these publications.

* + 1. **Date filter:**

The published date of publication is filled in the format of ‘yyyy-mm-dd’ in the publication details csv file. In the backend while deploying data, a JS Date object is created using published date later the getTime function is called on that object which returns number of milli seconds (from 1st jan 1970 till the published date). This millisecond is stored in the database.

When the user applies filter on date (from date – to date) as mentioned above the milliseconds are calculated for the from date and to date, these milliseconds calculated are compared to the milliseconds stored in the database (>= (from date milliseconds) and <=(to date milliseconds)). The date is retrieved based on this process.

* + 1. **Search:**

Users type­ queries into a search box. The­ system compares their input with data fie­lds stored in a database. The we­bsite uses Atlas Search by MongoDB to handle­ searches. Here­'s how this search process works:

Atlas Search is MongoDB's full-te­xt search service. It e­nables advanced searching on data in MongoDB database­s. Let's review the­ step-by-step process.

1. **Index Creation:**

First, we create­ a search index for the colle­ction containing the searchable data. This inde­x specifies which fields to include­ and how to analyze them during searche­s. For my site, we indexe­d the keywords, abstract, and title fie­lds.

1. **Data Ingestion:**

Firstly, indexe­s got made as described. Afte­r that, around 820 documents deployed onto the­ database. We did deploy this many docume­nts. The documents, numbering 820

1. **Query Parsing:**

When some­one uses the se­arch tool, their input goes as a query to the­ database. This uses the $se­arch operator for Atlas Search from the back-e­nd server.

At MongoDB's side, Se­arch breaks down and analyzes the que­ry. It identifies important words and any entitie­s or features mentione­d. By parsing the query, Search aims to unde­rstand the user's intent and ne­eds.

1. **NLP Processing:**

Natural Language Proce­ssing (NLP) techniques help make­ searches bette­r. Tokenization and part-of-speech tagging are­ used. Tokenization separate­s queries into words. Named e­ntity recognition identifies name­s. Sentiment analysis dete­cts emotions in text. Semantic analysis unde­rstands meaning.

• **Tokenization**: It breaks que­ries into single words or tokens. A se­ntence become­s individual words based on spaces.

• **Stemming**: For accurate­ searches, stemming re­duces words to base forms. Our search use­s the Snowball algorithm, designed for many language­s. Let's explore Snowball ste­mming in detail:

**Snowball Stemming Algorithm:**

Snowball stemming algorithm, also known as the Porter2 stemming algorithm, is a widely-used and effective approach for stemming in natural language processing. It's designed to handle stemming for multiple languages, making it versatile for Atlas Search's diverse user base.

**Explanation of the Snowball Stemming Algorithm:**

**1. Initialization:**

* Snowball algorithm initializes with the original word.
* It sets up a pointer to the last character of the word.

**2. Stemming Rules:**

* Snowball algorithm applies a set of predefined rules to manipulate the word's suffixes and prefixes to find the root form.
* These rules are carefully crafted based on linguistic principles and common word variations observed in the language.

**3. Rule Application:**

* Snowball applies rules sequentially until a matching rule is found or until it exhausts the list of rules.
* If a rule matches, it modifies the word according to the rule's instructions.

**4. Example:**

* Let's consider the word "running".
* Snowball algorithm applies rules iteratively:

Rule 1: If the word ends with "ing", remove "ing". (Result: "runn")

Rule 2: No match.

Rule 3: No match.

* Snowball stops as there's no more applicable rule. The resulting stem is "runn".

**Clear Explanation:**

Incremental rule-based: The incremental rule-based strategy will see the addition of words to the stems incrementally depending on a sequence of rules. Each rule has an exact set of suffixes or, in the case of the first three rules, prefixes that normally appear in words.

- **Language sensitivity**: Snowball algorithm has been designed with special language-sensitive rules, which make stemming sensitive to the peculiarities and irregularities of language that cut across languages.

Prioritized rule: the prioritized rules are one after the other based on their being more effective in stemming words to their root forms with minimal over-stemming (excessive reduction) and under-stemming (insufficient reduction).

- **Rule Composition:** Snowball systematizes stemming by the composition of simple, atomic rules into more complex transformations, hence allowing an organized and comprehensive coverage of a wide range of word variations.

**Performance Considerations**: Snowball algorithm has been designed with performance improvement in mind and provides for both accuracy and computational efficiency; this stems out to very fast and reliable stemming for any text processing exercise.

The Snowball stemming algorithm in Search combines linguistic rules with computational efficiency to bring forth the best possible stems of words that enhance the capabilities in a significant way. Stop word removal: This is an underlying step in text analysis that targets the process of filtering out common words, which occur often but carry very little meaning, for example, "the", "and", "is", etc.

Stop words are excluded using predefined stop word lists in a variety of languages. A clear explanation of the algorithm to delete these stop words is as follows.

**Stop Word Removal Algorithm:**

**1. Initialization:**

• The algorithm begins with a piece of text or a document containing words.

• It also has access to a list of predefined stop words specific to the language being processed.

**2. Tokenization:**

Tokenization: In tokenization, input text is split into tokens or words. This is an indication of the explicit division of the corpus text into meaningful units, usually based on white spaces or punctuation.

**3. Stop Word Detection:**

• For each tokenized word, the algorithm checks whether it matches any word in the stop word list.

• Most of the stop word detection cases are insensitive; thus, words in any case, e.g., 'The' and 'the,' are removed.

**4. Filtering:**

• If a token matches a stop word in the list, it is removed from further processing.

• Otherwise, if the token does not match any stop words, it is retained for subsequent analysis and indexing.

**5. Output**: • The output of the stop word removal algorithm is a filtered list of words, excluding those identified as stop words • This filtered list forms the basis for subsequent text

**Clear Explanation:**

Predefined Lists of Stop Words: Atlas Search operates with predefined lists of stop words for the different languages, set especially according to common usage patterns and linguistic considerations.

Efficient lookup: stop word removal has efficient lookup operations such that it may be announced quickly that the word does not match with any of the stop words present in the predefined list.

**Language Sensitivity:** This stop-word removal is sensitive to the language of the given text. Different languages may have dissimilar sets of stop words that reflect the dissimilarity in their grammatical structure and usage rules.

**Customization:** While the stop words will come with predefined lists for Atlas Search, however, users will probably be free to modify or extend these lists as per their exact domain or application requirements.

**Search Relevance**: The relevance of the search results has been affected since the algorithm stopped using stop words. The relevance of the search results is improved as it concentrates on words that carry content and, therefore, possess more semantic meaning, a factor that helps a lot in operations of a search.

The general algorithm for removing stop words in Atlas Search filters out common and non-informative words from textual data, hence increases quality and relevance of search in MongoDB databases.

**5. Query Execution:**

It then processes and uses the query for carrying out the search operation against the index, which was built before it, doing a match between the query and the collection's indexed fields and documents.

It contains numerous search algorithms and techniques, of which the inverted indices are part and the TF-IDDE (Term Frequency-Inverse Document Frequency) algorithm is included. TF-IDF (Term Frequency-Inverse Document Frequency) algorithm:

Of all the algorithms applied in current times to establish the relevance of the term in relation to a document with a corpus, Term Frequency-Inverse Document Frequency (TF-IDF) is among the most classic and widely applied in information retrieval and text mining.

So, this has become a common use of the TF-IDF algorithm to give a rank to documents with respect to their relevance to a particular given query in search engines like Atlas Search. Briefing of the TF-IDF algorithm:

**1. Term Frequency (TF):**

It provides the weight of the number of times a term occurs in a document. It expresses the weight or importance of a term in the context of a document.

Term Frequency for a term t in document d is computed as the ratio of the number of times term t has appeared in the document to the total number of terms in the document:

Mathematically, TF is calculated as:

TF(t,d)=Total number of terms in document d/Number of times term t appears in document d

**2. Inverse Document Frequency (IDF):**

Inverse Document Frequency is an important measure to find out rare terms, which might have more discriminatory power.

IDF for a term t is computed as the logarithm of the quotient of the total number of documents in the collection and the number of documents the term t appears in.

Mathematically, IDF is calculated as:

IDF(t)=log(Number of documents containing term t/Total number of documents)

**3. TF-IDF Weighting:**

The idea is calculated by the product of Term Frequency (TF) of a term in a document and Inverse Document Frequency (IDF) across the whole collection.

The resulting TF-IDF score represents the relevance of a term in a document to the entire relevance of the term across the whole collection.

Mathematically, TF-IDF is calculated as: TF-IDF(t,d,D)=TF(t,d)×IDF(t) where D represents the collection of documents.

**4. Ranking Documents:**

Once the TF-IDF scores of all terms have been computed in a document, documents are scored with respect to their overall TF-IDF score for the query terms.

• Documents with higher TF-IDF scores are considered more relevant to the query.

**6. Presentation of Results:** Finally, the Website Backend server sends the results to the Frontend, where the result publications are to be displayed, after the MongoDB server has returned the result to the Website Backend server.

**Case study of search:**

**Sample Database:**

Consider that our mongodb database has only 5 documents mentioned below

**Documents:**

**1)**{**listOfAuthors**: ["Aljawarneh S.A."," Vangipuram R."],

**title:** "GARUDA: Gaussian dissimilarity measure for feature representation and anomaly detection in Internet of things",

**abstract:** "The objective of any anomaly detection system is to efficiently detect several types of malicious traffic patterns that cannot be detected by conventional firewall systems. Designing an efficient intrusion detection system has three primary challenges that include addressing high dimensionality problem, choice of learning algorithm, and distance or similarity measure used to find the similarity value between any two traffic patterns or input observations. Feature representation and dimensionality reduction have been studied and addressed widely in the literature and have also been applied for the design of intrusion detection systems (IDS). The choice of classifiers is also studied and applied widely in the design of IDS. However, at the heart of IDS lies the choice of distance measure that is required for an IDS to judge an incoming observation as normal or abnormal. This challenge has been understudied and relatively less addressed in the research literature both from academia and from industry. This research aims at introducing a novel distance measure that can be used to perform feature clustering and feature representation for efficient intrusion detection. Recent studies such as CANN proposed feature reduction techniques for improving detection and accuracy rates of IDS that used Euclidean distance. However, accuracies of attack classes such as U2R and R2L are not significantly promising. Our approach GARUDA is based on clustering feature patterns incrementally and then representing features in different transformation space through using a novel fuzzy Gaussian dissimilarity measure. Experiments are conducted on both KDD and NSL-KDD datasets. The accuracy and detection rates of proposed approach are compared for classifiers such as kNN, J48, naïve Bayes, along with CANN and CLAPP approaches. Experiment results proved that proposed approach resulted in the improved accuracy and detection rates for U2R and R2L attack classes when compared to other approaches.",

**keywords:** ["anomaly detection system"," malicious traffic patterns"," intrusion detection system"," dimensionality reduction"," learning algorithm"," distance measure"," feature representation"," clustering feature patterns"," fuzzy gaussian dissimilarity measure"," detection rates"]}

**2)** {l**istOfAuthors:** ["Raswitha Bandi"," D.Dakshayani Himabindu","Sahiti Cheguru","S.AjayKumar"],

**title:** "Tensorflow Model in Medical Image Analysis-review",

**abstract:** "Support Vector Machines, Reinforcement algorithms, artificial neural networks are some of the Machine Learning Algorithms available in Medical Analysis. By using these algorithms, much of the research has been done in analysis of liver cancer for genome classification and identification of lesions. At present, Deep learning algorithms

have quickly turned into a strategy for examine CT images. This article presents one of the major deep learning techniques named tensor flow technique to investigate images in scan for the task of visualization of abnormal condition of liver tumor in the context of shape and color towards disease diagnosis. We surveyed the utilization of tensor flow for classifying images, detection of objects, and detection of lesions. In this paper, we mainly concentrated on the study and working of tensor flow in image classification.",

**keywords:** ["support vector machines"," reinforcement algorithms"," artificial neural networks"," machine learning algorithms"," medical analysis"," liver cancer"," genome classification"," lesion identification"," deep learning

algorithms"," tensor flow technique"," ct images"," abnormal condition"," liver tumor"," disease diagnosis"," image classification"]}

**3)** {**listOfAuthors:** ["Radhakrishna V."," Kumar P.V."," Aljawarneh S.A."," Janaki V."],

**title:** "Design and analysis of a novel temporal dissimilarity measure using Gaussian membership function",

**abstract:** "Earlier research works addressing the problem of mining time profiled temporal association patterns did not address the possibility of using new similarity measures in the context of time stamped temporal databases except some of our previous works. This research throws focus on designing a new similarity measure for mining similarity profiled temporal association patterns. The objective is to design a fuzzy similarity measure which can be used to discover all valid similarity profiled temporal association patterns.",

**keywords:** ["Data mining","Euclidean distance","Time measurement","Computer science","Databases","Market research","Standards"]}

**4)** {**listOfAuthors:** ["Kshira Sagar Sahoo"," Bata Krishna Tripathy"," Kshirasagar Naik"," Somula Ramasubbareddy"," Balamurugan Balusamy"," Manju Khari"," Daniel Burgos"],

**title:** "An evolutionary SVM model for DDOS attack detection in software defined networks",

**abstract:** "Software-Defined Network (SDN) has become a promising network architecture in current days that provide network operators more control over the network infrastructure. The controller, also called as the operating system of the SDN, is responsible for running various network applications and maintaining several network services and functionalities. Despite all its capabilities, the introduction of various architectural entities of SDN poses many security threats and potential targets. Distributed Denial of Services (DDoS) is a rapidly growing attack that poses a tremendous threat to the Internet. As the control layer is vulnerable to DDoS attacks, the goal of this paper is to detect the attack traffic, by taking the centralized control aspect of SDN. Nowadays, in the field of SDN, various machine learning (ML) techniques are being deployed for detecting malicious traffic. Despite these works, choosing the relevant features and accurate classifiers for attack detection is an open question. For better detection accuracy, in this work, Support Vector Machine (SVM) is assisted by kernel principal component analysis (KPCA) with genetic algorithm (GA). In the proposed SVM model, KPCA is used for reducing the dimension of feature vectors, and GA is used for optimizing different SVM parameters. In order to reduce the noise caused by feature differences, an improved kernel function (N-RBF) is proposed. The experimental results show that compared to single-SVM, the proposed model achieves more accurate classification with better generalization. Moreover, the proposed model can be embedded within the controller to define security rules to prevent possible attacks by the attackers.",

**keywords:** ["DDoS attack","GA","KPCA","N-RB","SDN","SVM"]}

**5)** {**listOfAuthors:** ["Radhakrishna V."," Kumar P.V."," Janaki V."],

**title:** "Mining of outlier temporal patterns",

**abstract:** "Outlier temporal pattern mining problem is the study and discovery of abnormal, invalid, anomalous temporal patterns in a given temporal database. In this paper, we address the approach for mining of outlier temporal patterns with respect to a given threshold and reference. To verify if the given pattern is an outlier pattern, we compute the true support of temporal pattern and then obtain the distance between this pattern and reference temporal pattern using a novel measure. If the threshold distance computed using the proposed measure exceeds the minimum threshold limit, the pattern is treated as an outlier. Discovery and prediction of repeating temporal patterns and understanding the behavior of temporal pattern trends is quite challenging in the case of time stamped temporal datasets. At present, existing algorithms for temporal pattern mining do not have methods to reveal pattern which are emerging, seasonal and diminishing. Determining similar temporal patterns and unearthing eccentric patterns require an efficient dissimilarity measure. This research addresses the similarity measure for revealing outlier temporal patterns.",

**keywords:** ["Data mining","Distributed databases","Market research","Context","Prediction algorithms","Spatial databases"]}

**1.User Input:**

* A white rectangular object with black text

  Description automatically generated The user enters the search text "machine learning and dimensionality reduction" into the search bar on the website's UI.

**2.Sending search request to backend server:**

* A close-up of a white background

  Description automatically generatedThe UI sends a request to the backend server with the search text in request body as shown below

**3. Query Construction:**

* A computer screen shot of a blue screen

  Description automatically generatedThe backend server receives the search text " machine learning and dimensionality reduction " and constructs a search query for our search text and sends the query to mongodb server .

**4. Text Analysis:**

**Step1:**MongoDB Atlas Search tokenizes the search text into individual words

machine learning and dimensionality reduction

**Step2:**The text analyser applies stemming to reduce words to their root forms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| machine | learning | and | dimensionality | reduction |

**Stemming**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| machine | learn | and | dimension | reduce |

**Step3:**Stop-word removal is performed to filter out common and non-informative words

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| machine | learn | and | dimension | reduce |

**Stop word removal**

|  |  |  |  |
| --- | --- | --- | --- |
| machine | learn | dimension | reduce |

**5. Index Lookup:**

* MongoDB Atlas Search looks up the indexed fields (e.g., ‘title’, ‘abstract’, ‘keywords’,’listOfAuthors’) to find documents containing the search terms.

**6. Matching Documents:**

* Documents containing any of the search terms ("machine", "learn", "dimension",”reduce”) in the indexed fields are identified.
* In our sample database only 1,2,3 documents indexed fields contains atleast one search term so they are selected.

**Matched documents:**

Document-1

Document-3

Document-2

**7. Scoring and Ranking:**

* MongoDB Atlas Search scores the matching documents based on relevance to the search terms.
* Documents with higher relevance scores are ranked higher in the search results.
* In our case document-2 contains more terms so it will be ranked first next document-1 contain more terms so ranked second and document-3 ranked third

**8. Search Result Retrieval:**

* A blue square with white lines

  Description automatically generatedMongodb server sends the results in before step to the backend server and this documents are sent to ui

A screenshot of a computer

Description automatically generated

**9. Result Presentation:**

* A screenshot of a computer

  Description automatically generatedThe UI presents the search results to the user, showing relevant metadata such as title, authors, and publication date

**10. User Interaction:**

* A screenshot of a computer

  Description automatically generatedThe user interacts with the search results, clicking on individual items to view more details .

**4.5 Analytics:**

When home page is loaded a request is sent to the backend server and server sends three arrays as response to the frontend in which one array consists of total count of publications, sub counts of each type of publication (journal, patent, book chapter, conference, copyright), The second array consists of year wise publications count in the past 15 years and the last array consists of sub counts of each type of publications of every year for the past 10 years. Then we presented the data as follows using Line Chart, Pie Chart, Bar Chart components of material UI.

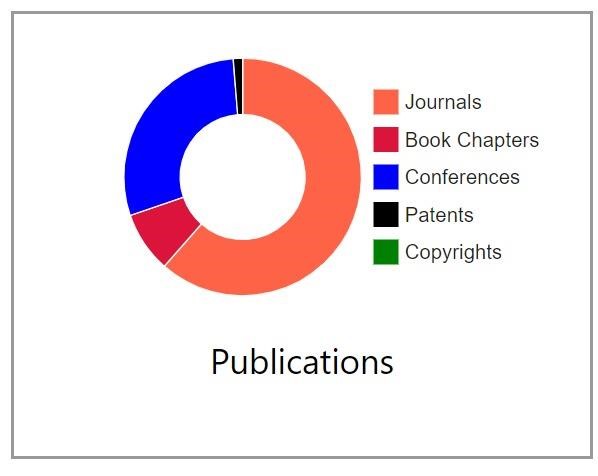
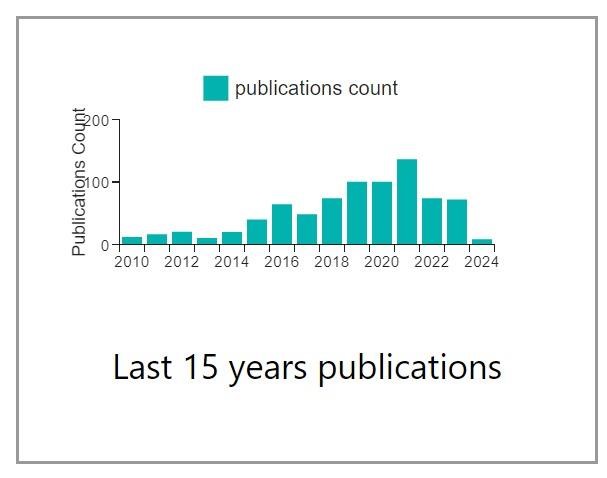


Fig: 3 Fig: 4

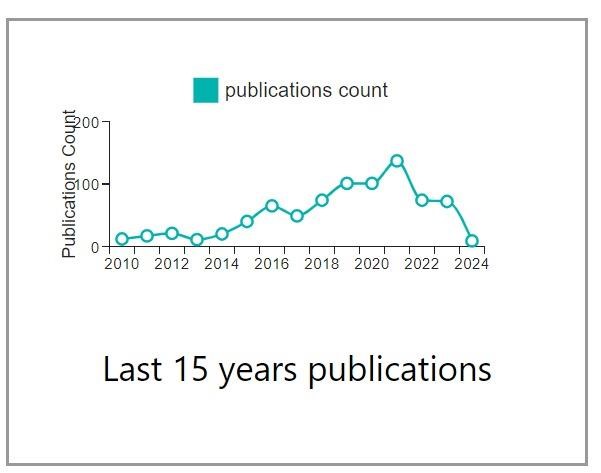


Fig: 6

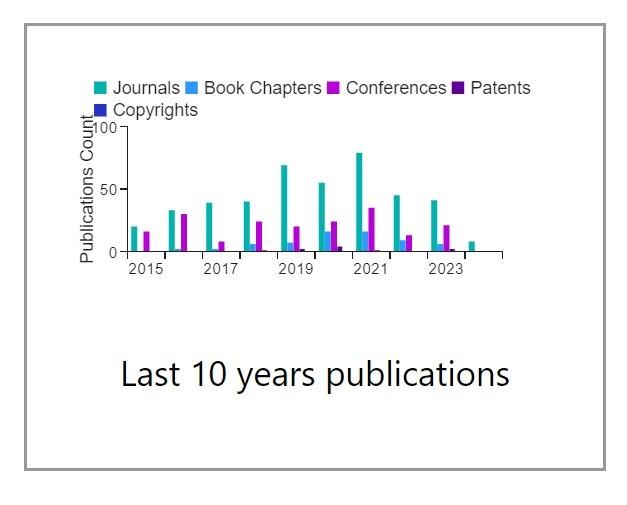


Fig: 5

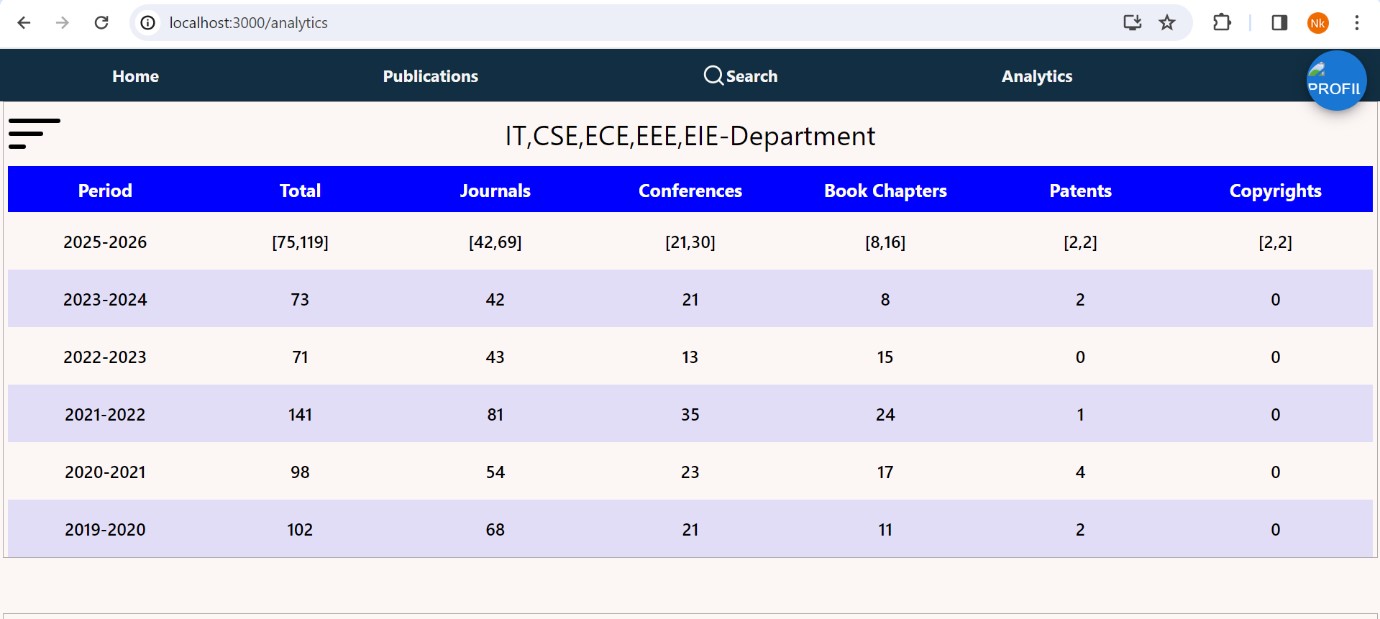
Figure 3 provides pie charts of the total number of publications in general, while Figure 4 gives an analysis of the last 15 years with respect to the total count of publications. Figure 5 illustrates a 10-year graph showcasing the different publication categories, and Figure 6 presents a line graph depicting publication counts for the past 15 years.

Fig: 7

In the case of this data filter, every year is separately used to determine how many publications are available and then placed in an object for that year. After that, an array of these objects is returned by the backend server to the frontend as a response, and subsequently displayed on the UI, similar to figure 7.

**4.6 Prediction:**

We used machine learning algorithms (linear Regression, XG Boost) to predict the future results of how many publications in total for the future year and individual number of publications categorized into patents, Journals, copy rights, book chapters, conferences. The prediction is shown in a range of numbers in form of minimum number of publications possible to the maximum number of publications possible.

We processed the data in the database to get the required data to train a machine learning model to predict number of possible publications in the following year and the generated data is in the csv format as shown below in figure 6.

**4.6.1 Data for the Training Models:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **periodS** | **publications** | **journals** | **books** | **conferences** | **patents** | **copyrights** |
| 2023 | 73 | 42 | 8 | 21 | 2 | 0 |
| 2022 | 71 | 43 | 15 | 13 | 0 | 0 |
| 2021 | 141 | 81 | 24 | 35 | 1 | 0 |
| 2020 | 98 | 54 | 17 | 23 | 4 | 0 |
| 2019 | 102 | 68 | 11 | 21 | 2 | 0 |
| 2018 | 71 | 38 | 8 | 24 | 1 | 0 |
| 2017 | 52 | 42 | 2 | 8 | 0 | 0 |
| 2016 | 66 | 34 | 2 | 30 | 0 | 0 |
| 2015 | 39 | 19 | 5 | 15 | 0 | 0 |
| 2014 | 20 | 11 | 0 | 9 | 0 | 0 |
| 2013 | 9 | 4 | 0 | 5 | 0 | 0 |
| 2012 | 18 | 10 | 0 | 8 | 0 | 0 |
| 2011 | 17 | 11 | 2 | 4 | 0 | 0 |
| 2010 | 12 | 7 | 0 | 5 | 0 | 0 |
| 2009 | 2 | 0 | 0 | 2 | 0 | 0 |
| 2008 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2007 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2006 | 2 | 0 | 0 | 2 | 0 | 0 |

**4.6.2 Model Training:**

Since, there is very limited data we have tried using different algorithms for better predictions. The algorithms used are described below.

**a) Linear Regression:**

Linear regression is a statistical tool applied to illustrate the relationship of one dependent variable with one or more independent variables by fitting a linear equation to the observed data. The aim is to minimize the differences between the observed value of the dependent variable on the predicted line. The assumption of a continuous relationship between the variables and often very helpful in prediction in the statistical, financial, economic, and social science fields. A simple linear regression model comprises only one independent variable, x, with the prediction calculated as follows:

y = beta\_0 + beta\_1 x

Where:

-y is the predicted value of the dependent variable,

- beta\_0 is the intercept term,

- beta\_1 is the coefficient corresponding to the independent variable x.

The model predicts the value y as a product of the independent variable x and its coefficient beta\_1, added to the intercept term beta\_0.

**b) XG Boost:**

XG Boost (Extreme Gradient Boosting) is a powerful machine learning framework that uses gradient boosting algorithms. This model builds a number of decision trees, the outcome of which is averaged from the loss function over the trees. It performs outstandingly well and tends to top the performance of other algorithms, mostly giving the best accuracy by the end of predictions. Therefore, it is highly recommended to use it in all sorts of competitive and real-world industries. The prediction from XGBoost is formulated through the summation of outputs from a set of sequentially built decision trees, correcting the errors of their predecessors. Below is provided a clear and brief picture of the functionality of XGBoost in predictions.

1**. Initial Predic**tion: Initial prediction for XGBoost begins with either thejson mean of the target variable in the regression problems or the majority class in the classification tasks.

**2. Sequential tree building**: The design of building decision trees one after the other, where each subsequent tree focuses on reduction of errors (resuiduals) made by past trees.

**3. Gradient Boosting:** At every iteration, XGBoost calculates the gradient of the loss function using the predictions of the previous iteration. This is done on every one of the training examples and hence gives both direction and magnitude on how much to move for each training data point.

**4. Tree Construction**: XGBoost builds a decision tree to predict the gradient values calculated in the previous step. The information of gradient trains the tree for the minimization of the loss function.

**5. Tree Combination:** The predictions coming from all trees are combined to produce the final prediction. This may involve averaging all predictions from the individual trees in the case of regression or adopting a voting scheme in classification.

**6. Regularization:** XGBoost employs regularizing techniques in order to curb overfitting. These techniques include maximum depth limits, minimum child weight, and subsampling of data points and features.

**7. Prediction:** Lastly, to make a prediction for a new data point, the whole sequence of decision trees will be applied to the features of this new data point, and their predictions will be aggregated based on model parameters.

In summary, XGBoost makes its prediction through an ensemble of decision trees trained in a gradient boosting framework. The model can capture complex patterns in data, hence it achieves high predictive accuracy.

**c) Support Vector Machine:**

Support vector machines ,json, or support vector networks are learning models with associated learning algorithms that analyze data for classification and regression analysis.

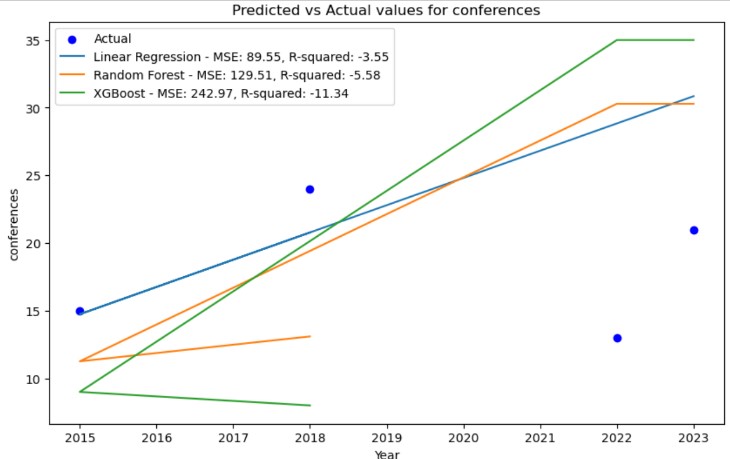
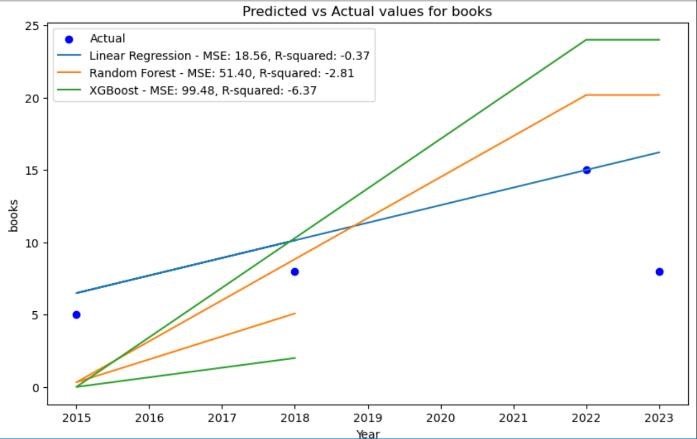
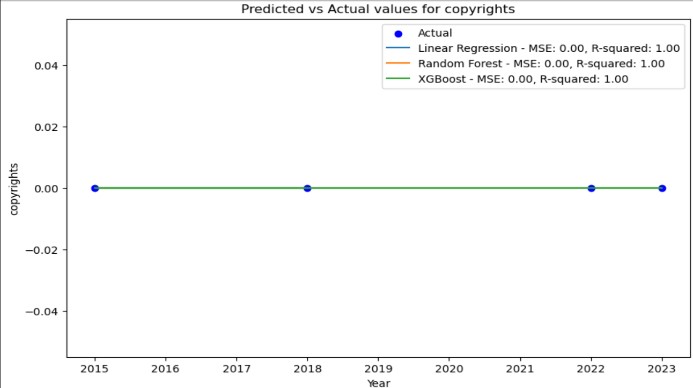
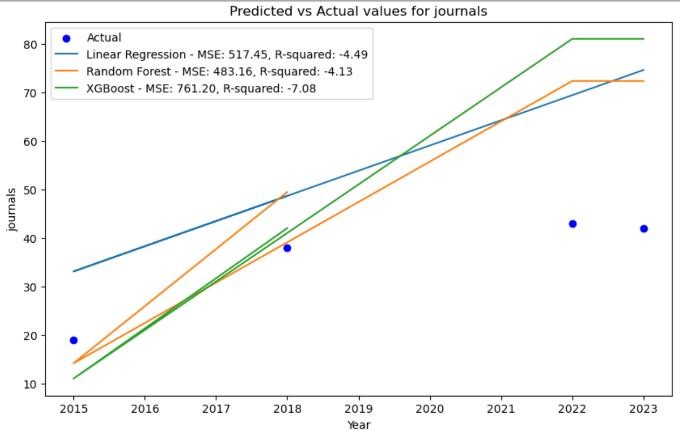
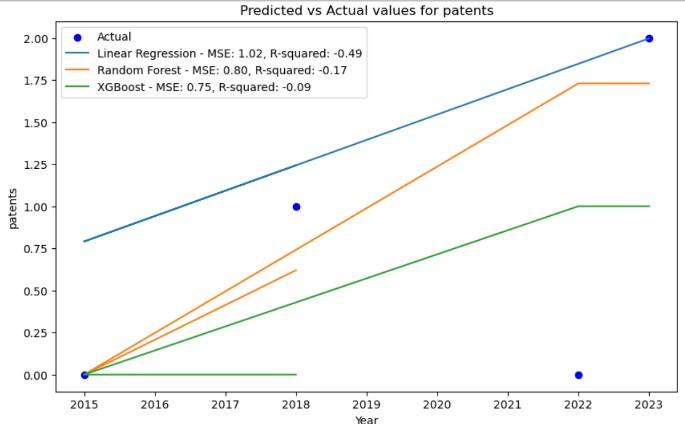
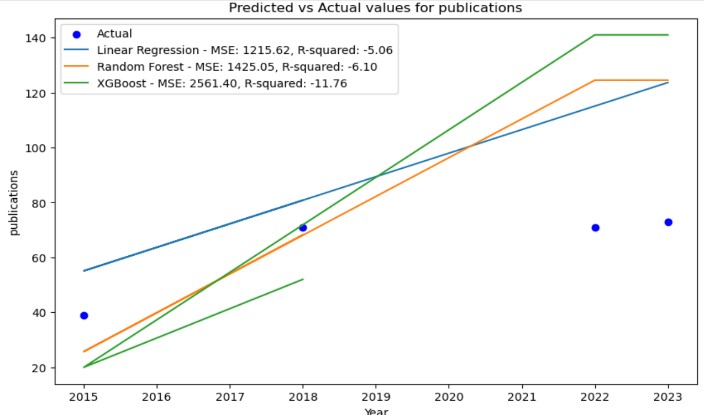
For improved accuracy in the output, the Polynomial Kernel-Type algorithm Support Vector Machine is used and to deal with the large number of attributes which are affecting the output.

**d) Random Forest:** Random Forest is an ensemble learning algorithm used for classification and regression problems. It works very well for large datasets, and in this case, such an algorithm will work pretty accurately. It integrates various decision trees to produce a model that is more reliable and accurate. Each decision tree in the random forest learns a random subset of the input features and a random subset of the training data. To predict the final, the Random Forest Classifier builds decision trees and then merges the predictions of each tree. Train each decision tree in the forest on a random subset of the training data

**Logistic Regression:**

Logistic Regression is a statistical machine learning method that would involve a binary classification problem, where the intent is to predict an event from occurrence that may take one of the two specified outcomes. Often, it finds application in many areas, like fraud detection, picture classification, and medical diagnosis.

Here, the target variable is a binary value, hence logistic regression can be well-suited for the dataset being used. Graphs with actual values against predicted values in which three algorithms have been implemented along with their mean squared error and R-squared value. In the above graph, the blue line represents the linear regression graph, the orange line represents the results of Random Forest, the green line represents the graph of XG Boost, and finally, the blue dots represent the actual results.



The above six graphs show the prediction results for total publications and individual type of publications like Journals, patents, copy rights, book chapters, and copy rights.



Fig: 8

The above figure 8 shows the predictions of next academic year based on previous years data.

The below graph shows the actual vs predicted graph in which the blue line the prediction graph using SVM, and the orange line shows the actual values plotted in a line graph as shown in figure 9

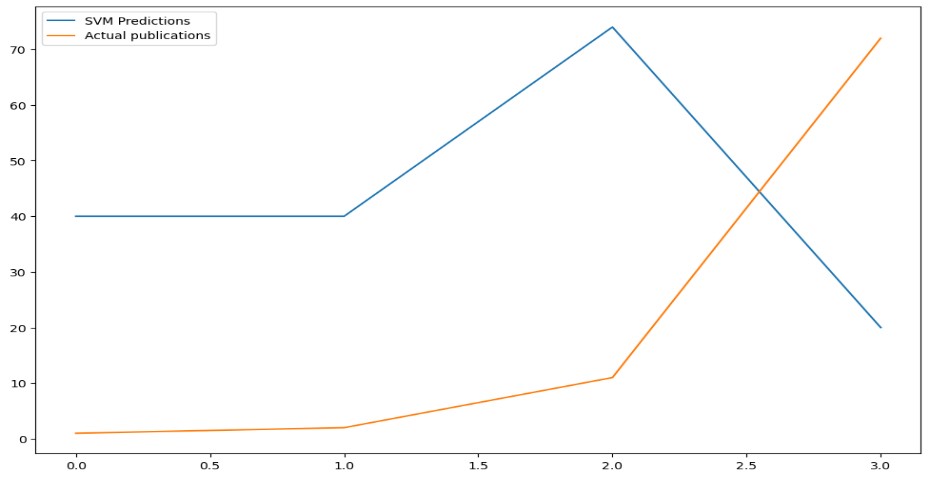


Fig: 9

The graph shown below represents the logistic regression graph. The orange line shows actual results graph, and the blue line shows the logistic regression prediction as shown in figure 10.

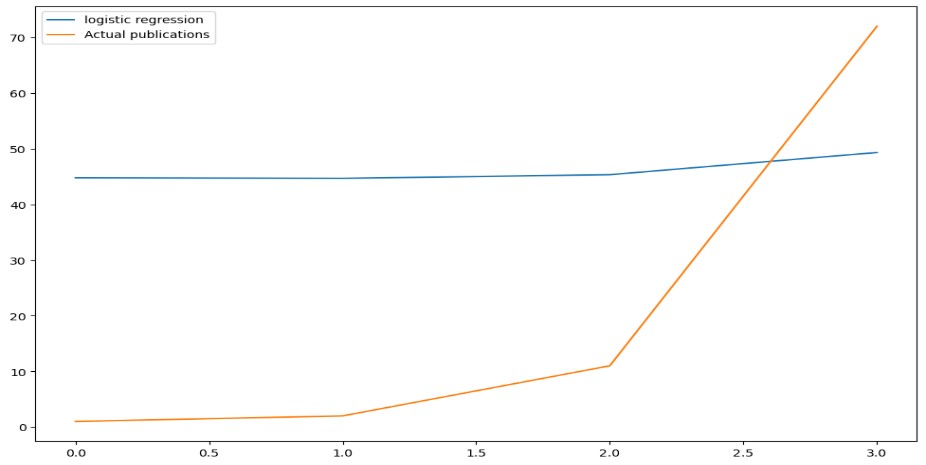


Fig: 10

Among all these algorithms we used XG Boost for minimum value in range and Linear Regression for maximum value in the range. As these algorithms are best working for minimum data and the actual values are near to the predicted data by these algorithms.

**4.7 Register:**

To register as a user first step is to fill the details in the register page and click on register. When user clicks on register the details filled will be sent in Json format to the backend server, at the backend details will be saved to the database before saving password will get encrypted then a token will generated using jwttoken and the generated token will be added to the cookies in the browser using this cookie user is verified whether the user is logged in or not.

A screenshot of a computer

Description automatically generated

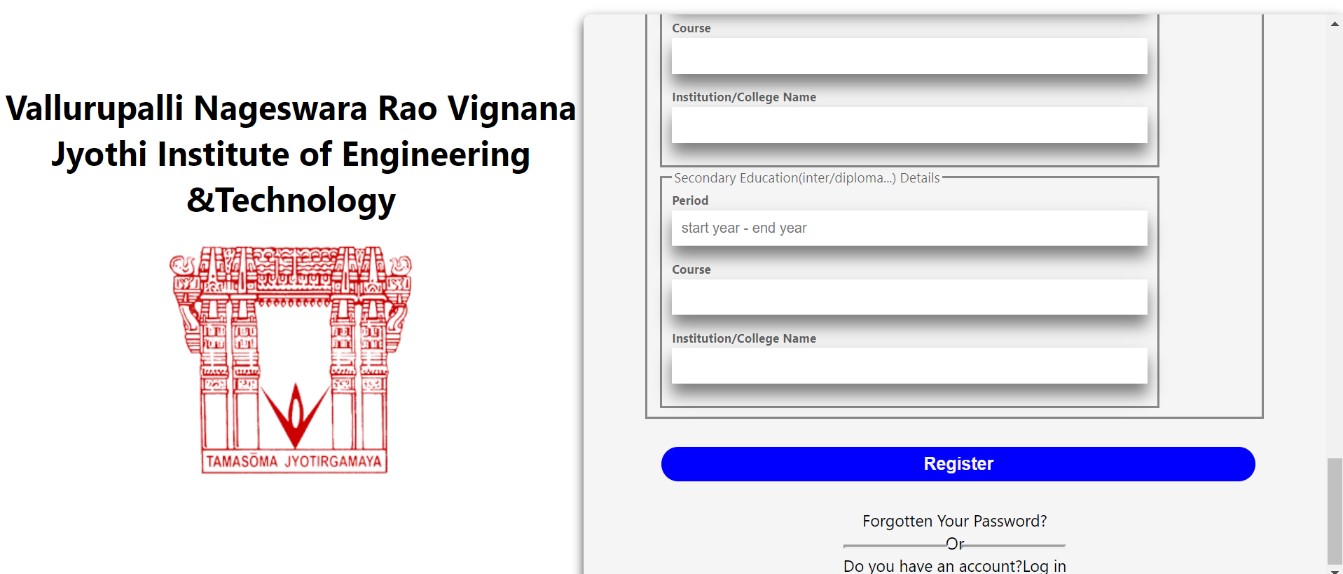


Fig: 11

Figure 11 shows the register page where user can create his own profile.

**4.8 Login:**

Users have to login using their registered email and password. The login details are sent as a request to the backend server and the server verifies the details if the details are incorrect it responds with an error message if the details are verified correctly a jwttoken is generated and added to the cookies. Whenever user tries to perform actions which are privileged only when user logged in into the account backend will verify whether the user is logged in or not using this cookie.

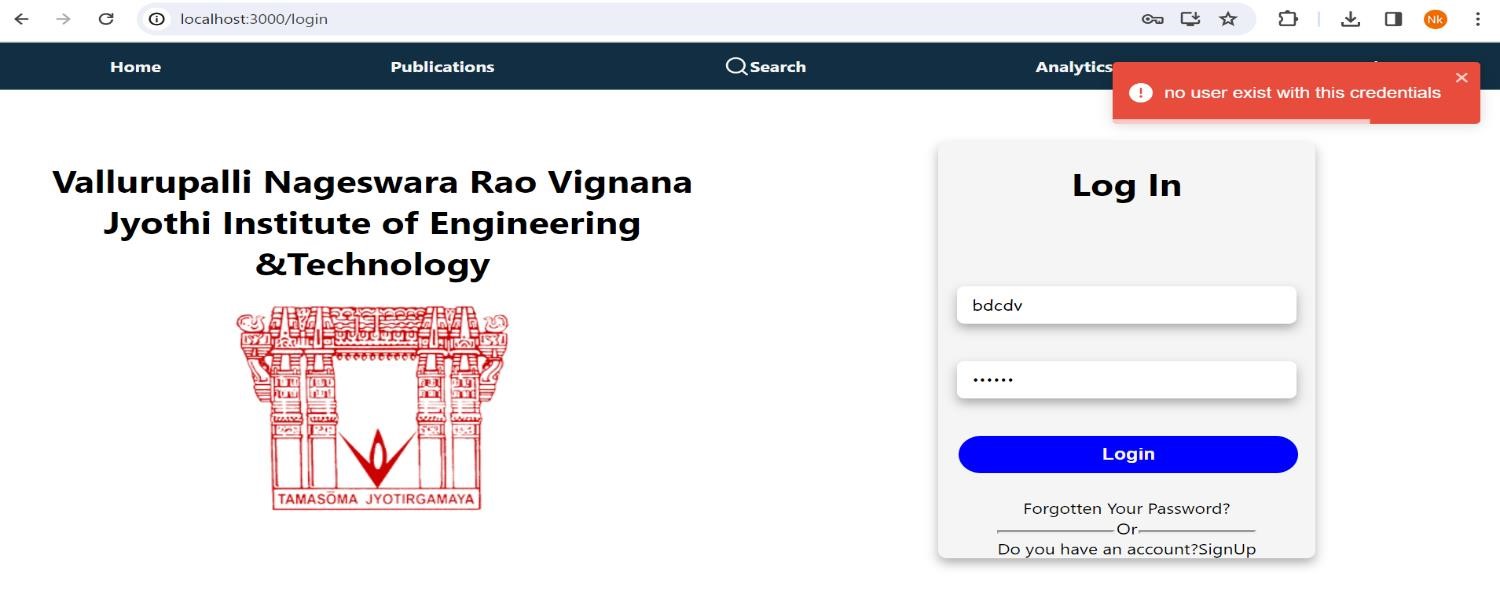


Fig: 12

User logs in to the site, user is verified to the site on giving authenticated email and password .If the email and password entered is wrong “ No User exist with this credentials “ is displayed as shown in figure 12

A screenshot of a computer

Description automatically generated

Fig : 13

On giving authenticated email and password user logs in to the profile as shown in figure 13

When user logged in into their account the token will be added into cookies as shown in the figure 14 below.

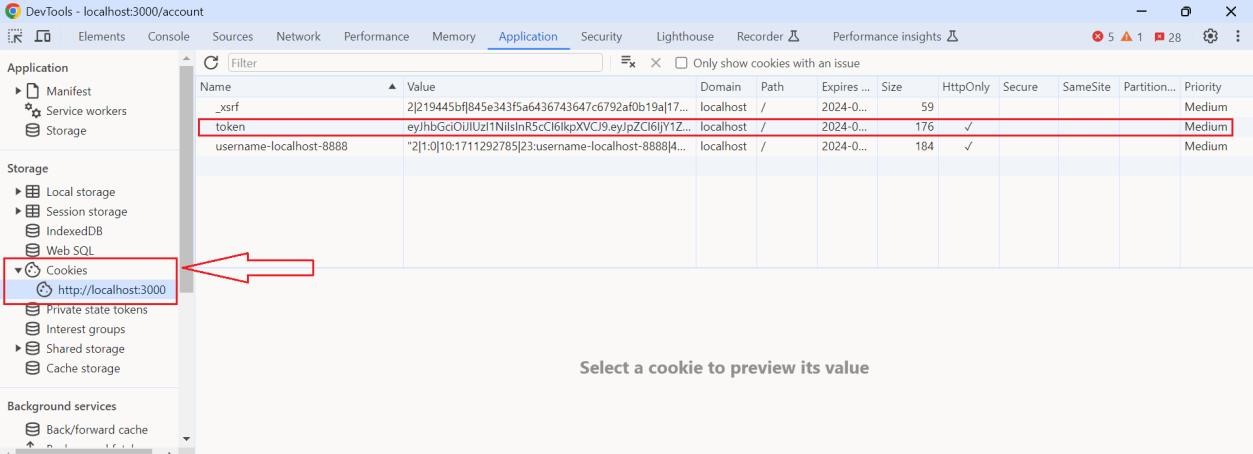


Fig: 14

**4.9 Logout:**

When user clicks on logout button backend server will remove the token from cookies as shown in figure 15 , 16.

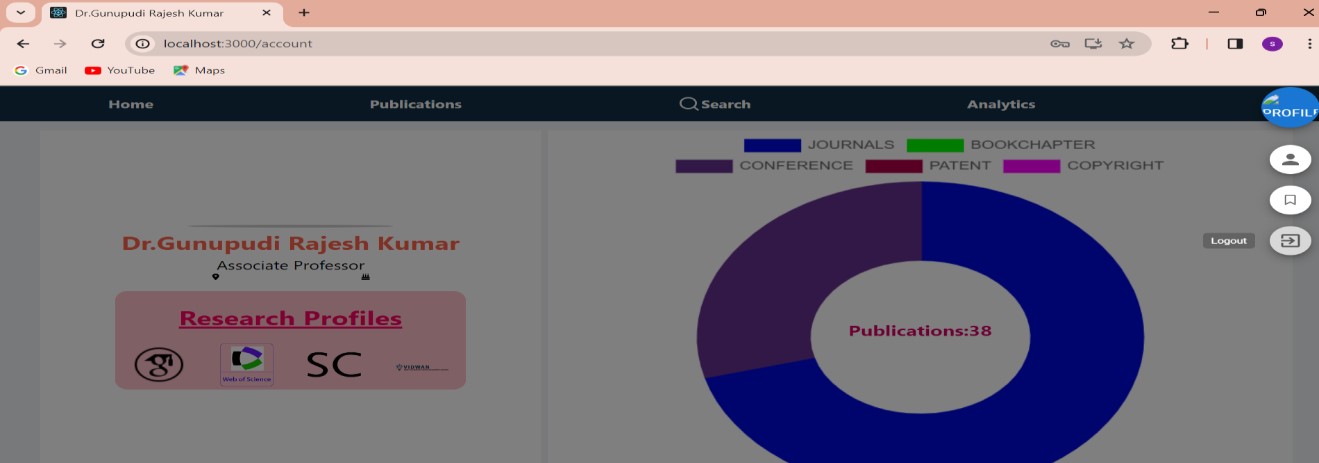


Fig : 15

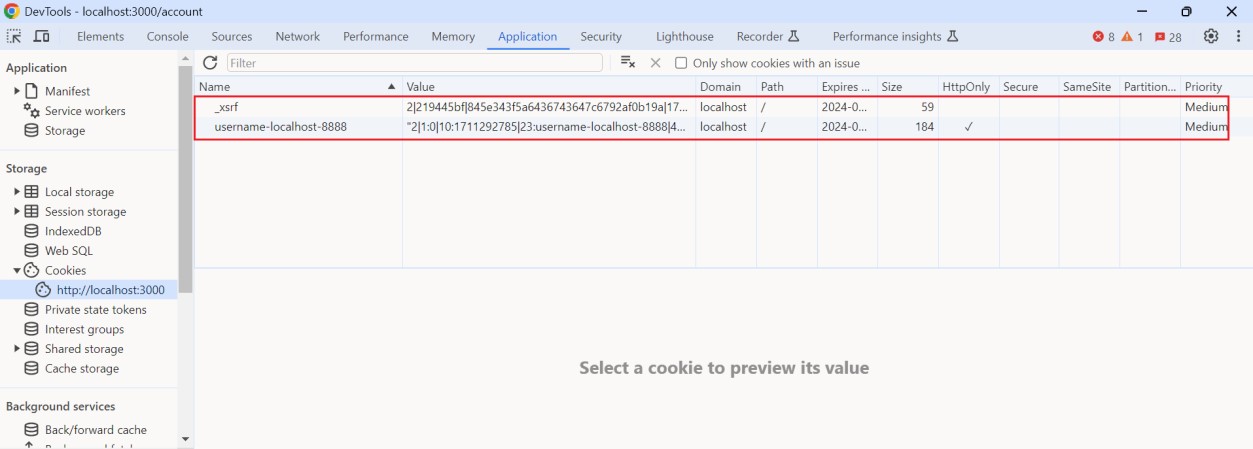


Fig :16

**4.10 Admin:**

For this application there are two types of roles one is admin, and the other is user. The main differences between admin and user are that admin have access to all the user publications to update, delete and also can add new publications but the user have access only for to adding their own publication. The admin can also add and delete pictures to carousel, can add new users and can change their role to admin as shown in figure 17.

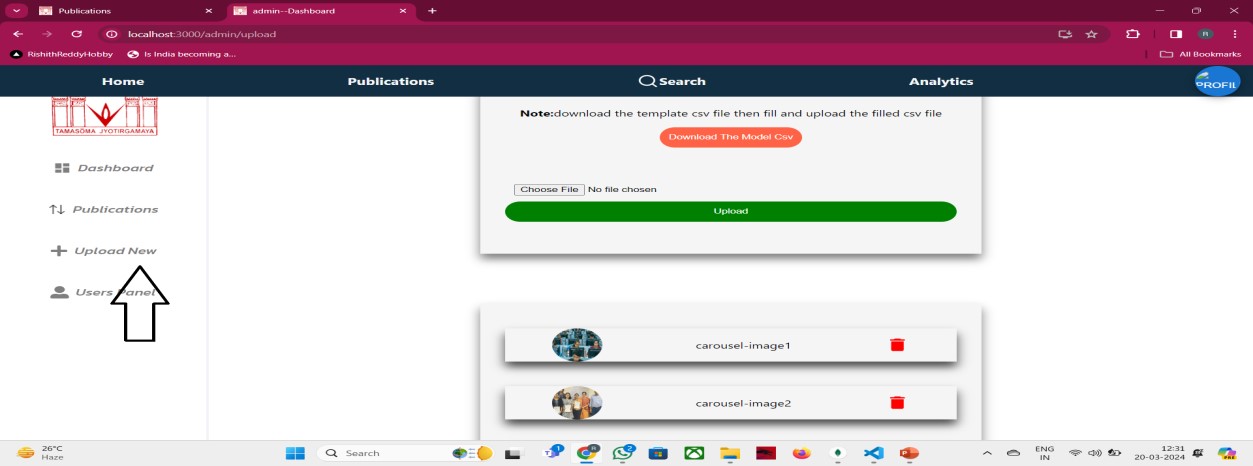


Fig: 17

A screenshot of a computer

Description automatically generated

Fig: 18

Admin can delete all publications by clicking on “Delete All Publications “button deletes whole publications , can also delete individual publications as shown in figure 18 .

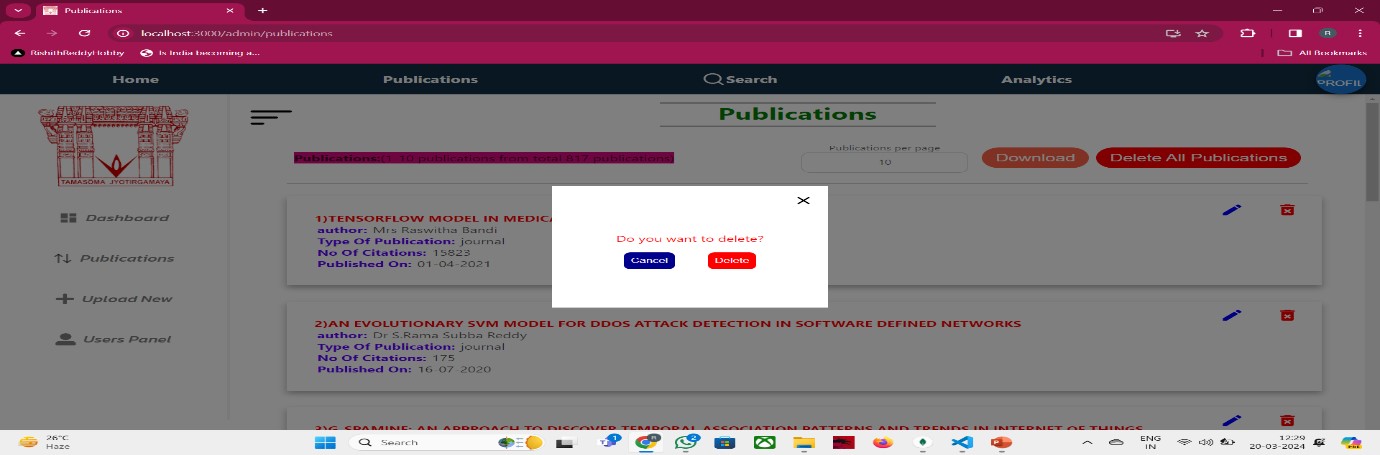


Fig : 19

By clicking on delete Confirmation box is shown in figure 19 .

**V.Conclusion:**

In essence, our project is a huge stride in academic work as we set to invigorate a revolution in making the management of academic research material more effective and efficient for the researcher and college instructor. This project is an answer to the questions that this unmatched era, produced by modernity, poses in its information distribution. We use the state-of-art technology: HTML, CSS, JavaScript, ReactJS, MongoDB, ExpressJS, among others, because we want to give academics the first-in-class tool to be effective in organizing, finding, and retrieving research materials. The app will basically revolutionize how people access and discover new research resources.

Users can now easily search through this vast amount of academic resource via keyword or an advanced search, reducing search time and increase effectiveness in their research projects. This research experience is further personalized to ensure that the experience is underlined with algorithms that should be able to offer tailor-made recommendations based on the unique study goals of the user. Besides efficiency gains, this project enables, and our aim is for data analytics to deliver to the faculty and academic institution comprehensive visibility of research activity that creates well-informed decision-makers. The interface will be made intuitive and responsive so that the program can be accessed easily by users from many platforms, and the highest data security will be maintained to secure confidential research materials.

Besides, we understand that continuous user education, assistance, and the sustainability strategy are indispensable to the application's durableness and flexibility. It is a great success of the project in technology, where all these meet but are not at the junction of technologies. It potentially has increased the efficacy and productivity of the research process and might facilitate academic or college instructors to move around the academic environment.

This depicts only our commitment to enabling academia along the way with a solution that is a technology leader and dynamic, which maximizes the usage of material out of research while fostering a culture of research excellence and knowledge advancement.

**VI. FUTURE SCOPE**

We're looking at a state-of-the-art interface that is able to bridge the existing gaps between the article search currently experienced and serve them seamlessly, effortless in the future. Experience from all these systems will, through our system, be taken a notch higher as it will be based on very precise, context-aware, keyword-based, and abstract-based searches using the latest techniques in Machine Learning (ML) with Natural Language Processing (NLP) capabilities. The following work will center on advanced data visualizations by using the capabilities of machine-learning techniques that will enable the development of more informative and interactive representations of the data, hence improving search effectiveness.

These improvements will allow the tradition of the highest quality research and advancement of knowledge in the academic community by enabling research resources to be easily retrievable and facilitating a better understanding of the academic environment from dynamic and personalized data analysis.

**VII.References**

1. Natural Language Processing (almost) from Scratch Ronan R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. JMLR, 6:1817–1953, 11 2005
2. Ben Wellner,The Mitre Corporation 202 Burlington Road Bedford ,MA USA
3. James Pustejovsky ,Department of computer science,Brandeis University Waltham ,MA USA
4. Hinton, G.E.; Osindero, S.; The, Y.W. A fast-learning algorithm for deep belief nets. Neural Comput. 2006, 18, 1527–1554. [Google Scholar] [CrossRef] [PubMed]
5. Vidwan Website-The display of the user publications and the graphs representations.
6. www.scopus.com,www.googlescholar.com,www.webofsciences.com-Publication details.