Artificial intelligence-enabled profiling of overlapping retinal disease distribution for ocular diagnosis

Sridhevi Sundararajan[1], Harikrishnan Ramachandran[2], Harshitha Gupta[2]
[1]Department of Engineering Design, Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, India
[2]Department of Electronics and Telecommunication Engineering, Symbiosis Institute of Technology, Symbiosis International Deemed University, Pune, India

Article Info

Article history:

Received Dec 16, 2023
Revised Feb 19, 2024
Accepted Feb 28, 2024

Keywords:

Comparative analysis
Disease frequencies and combinations
Ocular diseases
Retinal fundus
Retinal fundus multi-image dataset

ABSTRACT

Eyesight, an invaluable gift profoundly impacts our daily lives. In a rapidly evolving healthcare landscape, the preservation and enhancement of ocular health stand as critical objectives. This research endeavors to analyze the two retinal fundus multi-disease image datasets (RFMiD) one containing 3200 images and the other containing 860 fundus images. The primary objective of this study is to scrutinize these datasets, discern variations in the frequency of labeled diseases within and across them, and explore common combinations of labels. These findings hold important implications for the field of retinal image analysis, as they provide valuable insights into the distribution and co-occurrence of defects.

Corresponding Author:

Harikrishnan Ramachadran
Department of Electronics and Telecommunication Engineering, Symbiosis Institute of Technology
Symbiosis International Deemed University
Pune 412115, India
Email: harikrishnan.r@sitpune.edu.in

1. INTRODUCTION

In recent years medical image analysis domain has made notable progress with significant breakthroughs, revolutionizing the way healthcare professionals diagnose and manage various diseases. Among the areas of medical imaging, retinal disease analysis holds significant promise for improving the early detection, diagnosis, and treatment of a wide range of ocular conditions. Retinal diseases, including but not limited to age-related macular degeneration (AMD) [1], diabetic retinopathy (DR) [2], glaucoma [3], and retinal vein occlusion (RVO) [4], represent a substantial burden on global healthcare systems due to their prevalence and potential for vision impairment. In an era marked by the proliferation of data and its transformative potential across various domains, the analysis of datasets has become a cornerstone of research and decision-making processes. One such invaluable dataset, the retinal fundus multi-disease image datasets (RFMiD) [5], presents a unique opportunity to delve into the intricate details of retinal health and related diseases through multi-disease retinal fundus images [6]. RFMiD, comprising a diverse range of labels or attributes capturing essential information about retinal conditions, plays a pivotal role in understanding ocular health, disease progression, and treatment. As organizations and researchers increasingly harness the power of RFMID, the need for systematic approaches to analyze and compare these datasets has grown in importance. RFMiD [7] can vary significantly in terms of size, source, and labeling conventions, making it essential to establish robust methodologies for understanding their composition and characteristics. Such an understanding is crucial for optimizing diagnostic algorithms, treatment strategies, and furthering our knowledge of retinal diseases.

In this research endeavor, we delve into the analysis of multiple disease image datasets for the retina. Our primary objective is to gain comprehensive insights into the prevalence, associations, and patterns of retinal diseases across these datasets. To achieve this goal, we employ a data-driven approach, leveraging state-of-the-art data analysis tools and techniques. This research endeavors to address this need by presenting a comprehensive comparative analysis of two distinct RFMiD within the broader context of ophthalmic research. The primary objective is to unravel the label distribution and combinations within each dataset, shedding light on the relative importance of individual labels, such as disease types and the frequency in which they occur and the relationships between them. Notably, dataset A comprises 3,200 samples, while dataset B contains 860 samples, allowing for a detailed exploration of retinal health across a significant volume of data. Furthermore, this analysis aims to identify common label combinations that exist across the two datasets, revealing shared patterns and potentially cross-applicable insights in the field of ophthalmology. The analysis begins with the validation of data integrity, ensuring that label names are consistent between the two RFMiD. Subsequently, the study proceeds to calculate and report label frequencies and percentages within each dataset, providing a foundational understanding of label importance in the context of retinal health. Beyond individual labels, the research delves into the analysis of label combinations, exploring co-occurrence patterns of varying lengths and their relevance to disease diagnosis and prognosis. The results of this study would offer valuable insights for researchers, ophthalmologists, data scientists, and practitioners seeking to extract actionable knowledge from RFMiD. The generation of the dataset is depicted in Figure 1. By elucidating the distribution of labels and their combinations within the realm of retinal diseases, this research contributes to the growing body of knowledge in ophthalmic diagnosis, treatment, and patient care. In the following sections, we detail the methodology employed, present the findings, and discuss the implications of this comparative analysis, underscoring its significance for the broader field of ophthalmic research and data-driven healthcare.

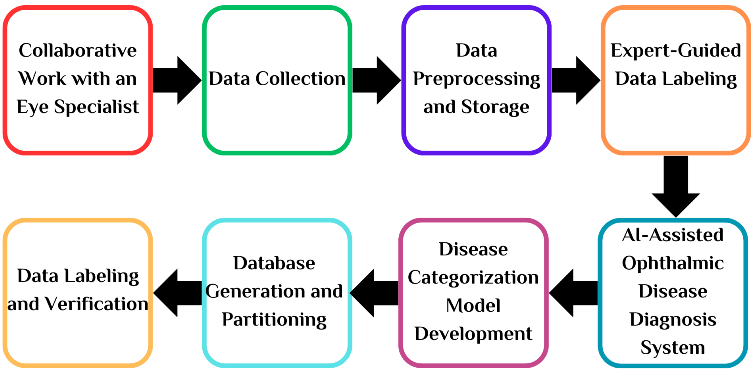
Collaborative Work with an Eye Specialist
Data Collection
Data Preprocessing and Storage
Expert-Guided Data Labeling
AI-Assisted Ophthalmic Disease Diagnosis System
Disease Categorization Model Development
Database Generation and Partitioning
Data Labeling and Verification

Figure 1. Creation of RFMiD approved from ophthalmologists in order to design artificial intelligence (AI) based disease diagnosis system [5]

Existing work on retinal fundal: the analysis of medical image datasets, particularly those related to ophthalmology, has gained significant attention in recent years due to its potential in aiding disease diagnosis, patient care, and medical research. In this section, we review the relevant literature that underscores the importance of comparative analysis of retinal image datasets. As well as the methodologies employed in similar studies in Table 1.

Challenges and opportunities: despite the promising advancements in the field, challenges persist in standardizing data collection, labeling, and analysis across ophthalmic datasets. Differences in imaging equipment, image quality, and labeling conventions necessitate careful comparative analyses to account for dataset-specific variations. Our research addresses these challenges by validating data consistency and exploring label distribution across two distinct RFMiD. In conclusion, the literature underscores the significance of comparative analysis in the context of ophthalmic image datasets. By leveraging the rich information within RFMiD and similar datasets, researchers and healthcare professionals can advance our understanding of retinal diseases and enhance patient care.

Table 1. Existing studies done on retinal fundal

Topic	Description
Optical coherence tomography	Optical coherence tomography [8] has emerged as one of the for runners in the technology used for retinal disease diagnosis as it provides high-resolution, cross-sectional images of retina therby enabling the detection of structural changes at a microsopic level.
AI [9] and machine learning for diagnosis [10]	Recent literature review has proved that AI and machine learning are highly dependable in automating the retinal disease diagnosis. Machine learning algorithms helps to analyze the retinal images with high accuracy thereby aiding in early detection of diseases such as DR, glucoma [11] and many other retinal diseases.
Utilization of ophthalmic image datasets [12]	Medical imaging [13], specifically ophthalmic imaging, has undergone a trans formative shift with the advent of digital technologies. High-resolution retinal fundus images, as exemplified by the RFMiD, provide a wealth of information for diagnosing retinal diseases, including DR, glaucoma, and AMD. Researchers and healthcare practitioners have harnessed these datasets to develop automated diagnostic systems [14], identify disease biomarkers [15], and assess treatment outcomes.
Comparative analysis in ophthalmic research [16]	Comparative analysis of retinal image datasets has become essential for understanding the variations in disease manifestations across diverse patient populations. Previous studies have successfully compared datasets of varying sizes and origins to elucidate the impact of demographic factors, such as age and gender, on disease prevalence and severity. Such analyses have led to tailored treatment strategies and improved patient care.
Label distribution and co-occurrence patterns [17]	Exploring label distribution and co-occurrence patterns within ophthalmic image datasets has emerged as a crucial research area. Our study builds upon this foundation by focusing on label combinations within the RFMiD.

Background of proposed work: the dataset used represents an insightful resource in the diagnosis of ocular diseases. RFMiD comprises retinal fundus images, capturing a range of eye conditions and diseases. These images are instrumental in the early detection, diagnosis, and monitoring of eye-related ailments, including DR, AMD, and glaucoma. The significance of RFMiD lies in its potential to revolutionize ophthalmological diagnostics and improve patient care. The dataset provides a diverse and comprehensive collection of retinal images, enabling researchers and healthcare professionals to develop machine learning and computer vision models for disease classification and severity assessment. However, the analysis of RFMiD involves multifaceted challenges, including image preprocessing, feature extraction [18], and disease classification [19]. The code logic developed for this research is to facilitate the exploration and analysis of the RFMiD, addressing the following objectives as mentioned in Table 2. By addressing these objectives, this code aims to contribute to the effective utilization of the RFMiD for medical research and diagnostics. It empowers researchers to assess dataset variations, understand disease distribution, and lay the groundwork for the development of advanced machine learning models for automated disease detection and analysis in retinal fundus images.

Table 2. Exploring retinal disease data: methodology, variations, and visualization techniques

Retinal disease dataset- study approach	Description
Data reading and preprocessing	Compare the two RFMiD comprising of 3200 and 860 images respectively to check for the Label[Disease Name] consistency in both the datasets.
Label frequency analysis	Perform frequency analysis to determine the occurance of the diseases and its percentage within each RFMiD respectively. This analysis will provide insights into the prevalance of specific diseases among the patients with ocular diseases.
Label combination analysis	Building on the insights from label frequency analysis, the research explores label combinations within the datasets, encompassing Identifying label combinations that occur within the data samples of each dataset. Quantifying the frequency of each label combination to comprehend its prevalence. Analyzing label combinations of varying lengths, such as pairs or triplets, to unveil co-occurrence patterns.
Data analysis	The comparative analysis of label frequencies and combinations was conducted for both RFMiD that includes creating graphical representations to illustrate label frequencies and co-occurrence patterns

2. METHOD

This section provides an comprehensive overview of the methodology utilized in our study, encompassing the research design, data set acquisition methods and data analysis techniques employed to achieve our research objectives. We present a detailed account of our approach, ensuring transparency and clarity in explaining how we conducted our study.

2.1. Setup and variables considered

2.1.1. Python environment

The code logic written for this study was developed in Python code using Python 3.10 version. The code written for this research study depends on essential libraries namely Pandas and Numpy for data manipulation and numerical operations respectively. The use of color fundus photos to visualize retinal circulation provides a valuable non-invasive method for examining the microcirculation within the human retina, allowing for a unique opportunity to assess systemic health. Thorough clinical analysis not only helps provide a detailed study about eye-ailments but also helps to detect certain chronic conditions namely diabetes stroke [20], hypertension, arteriosclerosis [21], cardiovascular diseases, neurodegenerative disorders, as well as renal and fatty liver diseases [22]. Hence screening of the eye together with timely consultation and treatment help in preventing not only the loss of vision but also helps in preventing any damage to other parts of the body. Some of the previous studies have come with datasets related to only a few diseases threatening the vision. However a need for multi disease dataset is felt in order to maintain a general retinal screening system. The data utilized in this study consists of two distinct datasets obtained from the RFMiD. These datasets, denoted as dataset A and dataset B, were subjected to a series of data preprocessing and analysis steps to fulfill the research objectives. The Table 3 describes the specifications of the datasets used in this research [7]. The RFMiD [7] and RFMiD2 [5] publicly accessible datasets include 3200 and 860 retinal images, respectively. Within the 3200 retinal images, 45 were identified as abnormal, signifying the presence of 45 distinct disease types in this dataset. Similarly, among the 860 retinal images, 49 were identified as abnormal.

Table 3. Retinal fundus data specification [5]

Subject area	Medical data in the field of ophthalmology
More specific subject area	Multiple disease classification of retinal fundus image
Category of data	Comma-seperated value files, images
Data acquisition	TOPCON TRC-NW300
Format of data	Tagging and annotating JPEG and PNG image files to create .CSV files
Variables under study	The majority of the patients received mydriasis through a single drop of tropicamide at a concentration of 0.5%. Non-mydriatic procedures were employed for certain participants.
Characteristics of the experiment	Fundus images were captured while the patient was in an upright position, with a distance of 40.7mm (TOPCON TRC-NW300) and 42mm (CARL ZEISS FF450) between the camera lenses and the eye under examination, employing a non-invasive fundus camera.
Location of data source	State of Art Eye Care Hospital known as Shri Ganpati Netralaya, situated in Jalna, Maharashtra, India, and the Center of Excellence in Signal and Image Processing, affiliated with SGGS Institute of Engineering and Technology, located in Nanded, Maharashtra, India.

2.2. Dataset processing

Based on these two data sets available we have arrived at an algorithm that gives an insight as mentioned in Figure 2. The analysis consists of several key steps. First, we compare the labels (disease attributes) between the two datasets. This involves extracting the label columns from both datasets, sorting them alphabetically for consistency, and identifying identical labels present in both datasets. Next, we calculate the frequency of labels within each dataset by summing up the values in their respective label columns. We also compute the percentage of each label within each dataset by dividing the label frequency by the total number of records and multiplying by 100. Additionally, the analysis involves determining label combinations within each dataset. To achieve this, the algorithm examines the label combinations in the dataset by iterating through the records and identifying labels with a value of 1 (indicating label presence). Combinations of these selected labels are generated using Python's itertools library, ranging from pairs to the total number of selected labels. The frequency of each label combination is calculated and stored for both datasets. The next step involves creating a table that records label combinations and their corresponding frequencies, sorted in descending order to identify the most common combinations. Finally, a table of common label combinations is generated by merging the combination label frequency tables for both datasets based on the label combinations with values greater than 0. This comprehensive approach provides insights into label variations and combinations within the datasets. CSV files that are described in the Table 4 are generated for the data processing methods. Overall 5 CSVs are generated based on the code logic applied using the datasets RFMiD and RFMiD2. The name of the CSVs generated depict the nature of the context present in them.

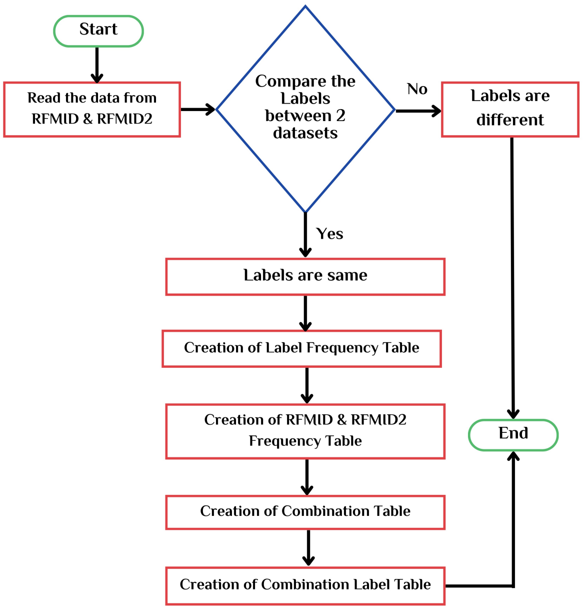
Start
Read the data from RFMID & RFMID2
Compare the Labels between 2 datasets
No
Labels are different
Yes
Labels are same
Creation of Label Frequency Table
Creation of RFMID & RFMID2 Frequency Table
Creation of Combination Table
Creation of Combination Label Table
End

Figure 2. Flowchart showcasing the methodology to determine the profiling of retinal disease distribution and overlapping patterns in multi-disease retinal fundus images for ocular patient

Table 4. Description of CSV files and their parameters

CSV file name	Parameters considered	Significance of the CSV
label-frequency-table.csv	Label, frequency RFMiD, percentage RFMiD, frequency RFMiD2, percentage RFMiD2	This table allows you to quickly compare label distributions between the two datasets. It reveals which labels are more prevalent and provides insights into the composition of each dataset.
RFMID-frequency-table.csv	Label-combination, RFMiD	It helps identify patterns and associations between labels in RFMiD dataset. Researchers and analysts can use this table to understand which combinations of attributes commonly appear together in the RFMiD dataset.
RFMID2-frequency-table.csv	Label-combination, RFMiD2	It helps identify patterns and associations between labels in RFMiD2 dataset. Researchers and analysts can use this table to understand which combinations of attributes commonly appear together in the RFMiD2 dataset.
labelcombination-table.csv	Label-combination, RFMiD,-RFMiD2	It provides a side-by-side comparison of label combinations between the two datasets. Researchers can use this table to identify commonalities and differences in label associations.
common-labeltable.csv	Label-combination, RFMiD,-RFMiD2	It highlights label combinations that are shared between the two datasets, which can be valuable for identifying consistent patterns across different versions or subsets of the data.

3. RESULTS AND DISCUSSION

Using the algorithm and the code that has been developed to study the datasets the results section describes the analysis of the study performed. Table 5 describes the abbreviations used across the study to describe the labels.

3.1. Label names consistency

Ensuring label name consistency is a important aspect of this study as it signifies that the two datasets utilized share a uniform set of labels. This consistency plays a crucial role for ensuring data compatibility across different datasets and facilitates accurate analysis. Without consistent label names, comparing and combining data from various sources becomes challenging, hindering the reliability and interpretability of the analysis results.

Table 5. Ophthalmology abbreviations description

Abbreviation	Full form
AION	Anterior ischemic optic neuropathy
AH	Asteroid hyalosis
ARMD	Age-related macular degeneration
BRVO	Branch retinal vein occlusion
DN	Drusens
DR	Diabetes retinopathy
ERM	Epiretinal membrane
EDN	Exudation
HR	Hemorrhagic retinopathy
HPED	Hemorrhagic pigment epithelial detachment
MCA	Macroaneurysm
MH	Macular hemorrhage
MYA	Myopia
ODC	Optic disc cupping
ODE	Optic disc edema
OCT	Optical coherence tomography
TSLN	Tortuous superficial large vessels

3.2. Label frequency analysis

The frequency analysis of the labels derived from the algorithm provides valuable information about the distribution of the labels in terms of their occurrences as well as their percentage in both the datasets.Notably, some labels exhibit consistent occurrence across both segments, such as ”DR” and ”EDN [23],” while others display notable disparities in frequency. The percentages provided offer a clear insight into the proportion of each label within its corresponding segment. This analysis sheds light on the distribution of labels within the dataset, serving as valuable information for further exploration or decision-making processes. In the datasets that are used in our study namely RFMiD and RFMiD2 the occurrence of DR is highest having a value of 632 in RFMiD and HR [24] is highest having a value of 86 in RFMiD2 respectively. Based on the comprehensive analysis, it is evident that DR is notably prevalent in both datasets. This finding underscores the importance of delving deeper into retinal research. Nevertheless, it is imperative to acknowledge the requirement for more balanced dataset distributions as an essential step for future investigations in the field of retinal health. The Figures 3 to 6 gives an overview about the frequency and percentages of the diseases for RFMiD and RFMiD2 datasets respectively.

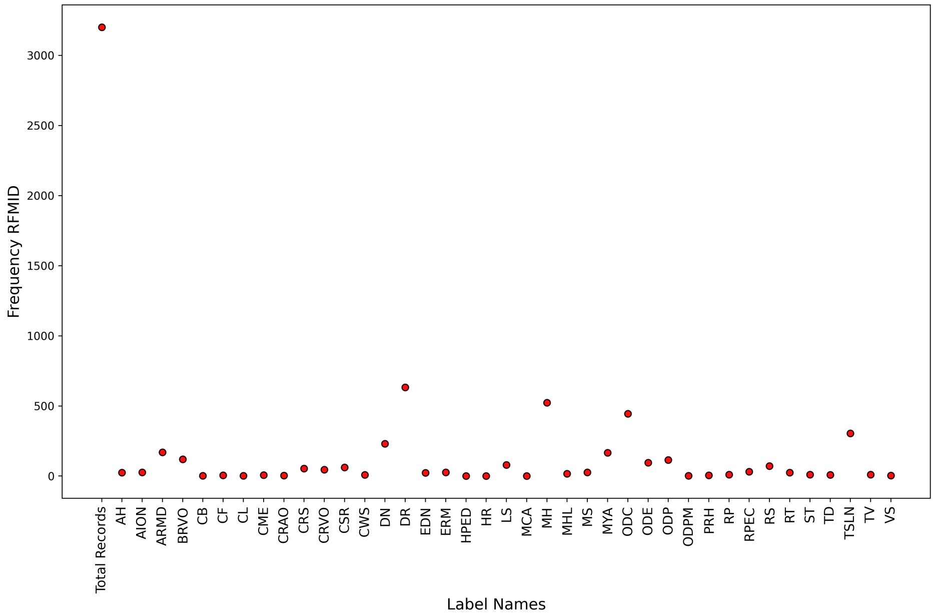

Figure 3. Scatter plot depicting the frequency distribution for RFMiD dataset

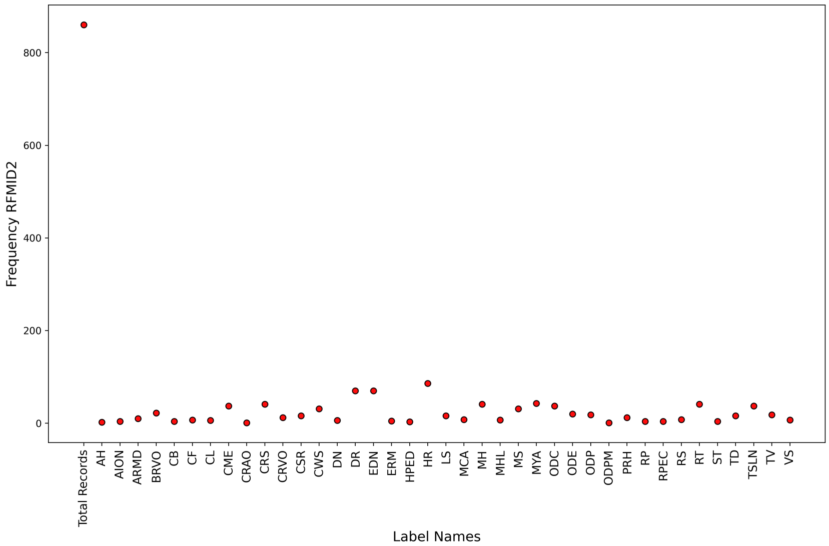

Figure 4. Scatter plot depicting the frequency distribution for RFMiD2 dataset

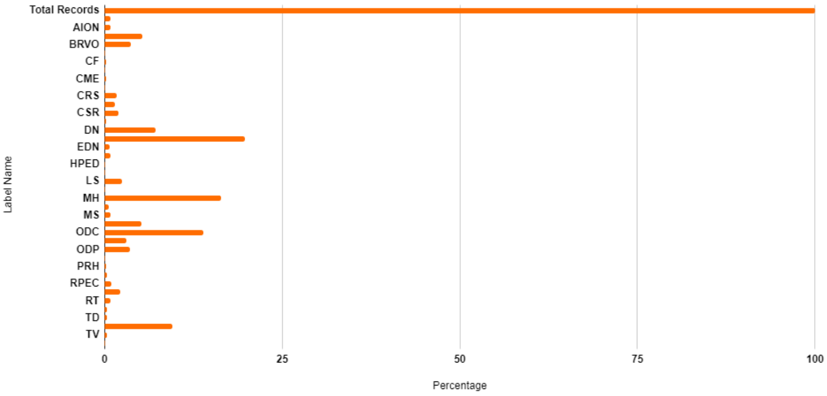

Figure 5. Bar chart depicting the percentage distribution for RFMiD dataset

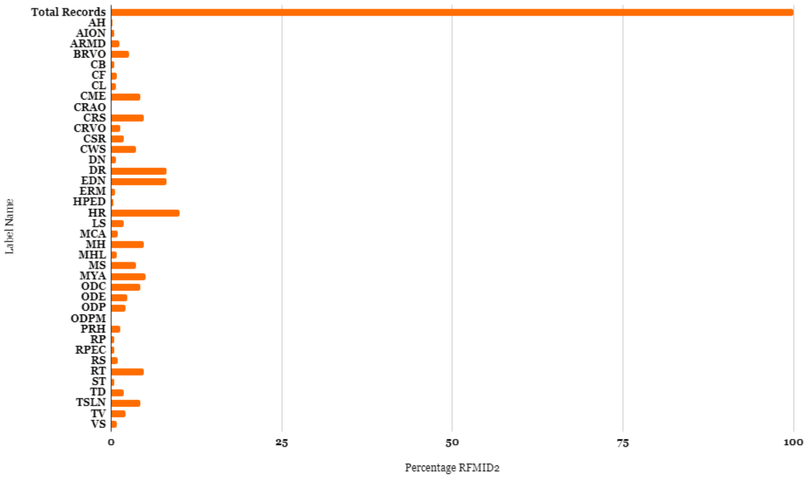

Figure 6. Bar chart depicting the percentage distribution for RFMiD2 dataset

3.3. Label combination analysis

The dataset reveals a diverse array of label combinations, reflecting the categorization of various entities. RFMiD and RFMiD2 values span from 0 to 83 and 0 to 9, respectively, with some combinations standing out with notably higher values, indicating their prominence. Among these combinations ODC and TSLN emerge as the most prevalent, boasting RFMiD and RFMiD2 values of 83 and 4, respectively. Other common pairings include ('MH', [25] 'ODC' [26]), ('MH', 'TSLN' [27]), ('DR', 'ODC'), ('DR', 'LS'), and ('DR', 'TSLN'). Certain labels such as 'ODC', 'TSLN', 'MH', and 'DR' appear frequently across various combinations, indicating their significance, while others like 'CB', 'ARMD', 'AION', and 'BRVO' [28] are less prevalent. The dataset demonstrates significant variability in RFMiD and RFMiD2 values across different combinations, reflecting diverse levels of importance or relevance. Combinations involving 'ODC', 'TSLN', 'MH', and 'DR' tend to exhibit higher values, suggesting their increased commonality or significance. Notably, combinations with multiple labels, particularly involving 'ODC', 'TSLN', and 'MH', often possess elevated RFMiD2 values, potentially indicating specialized subcategories within broader combinations. Trends within the dataset indicate that combinations featuring 'ODC', 'TSLN', 'MH', and 'DR' tend to have higher RFMiD and RFMiD2 values, implying their greater prevalence or importance. Moreover, certain combinations with multiple labels, especially those involving 'ODC', 'TSLN', and 'MH', exhibit notable RFMID2 values, suggesting potential subcategories or specialized groupings.

- Frequent combinations: frequent combinations, which serve as the backbone of the dataset, showcase patterns and relationships occurring with notable regularity. The dataset reveals several frequent combinations that shed light on prevalent patterns and relationships among the variables. For instance, 'ODC' and 'TSLN' emerge as one of the most frequent combinations, appearing with a high RFMiD of 83, indicating a strong association between these two factors. Similarly, the combination of 'MH' and 'ODC' holds significant prominence with an RFMiD of 77, suggesting a common occurrence of these variables together. Additionally, 'MH' and 'TSLN' exhibit a robust association, reflected in their RFMiD of 76. Moreover, 'DR' and 'ODC' form another notable frequent combination with an RFMiD of 72, indicating a recurring relationship between these variables. Lastly, 'DR' and 'LS' stand out with an RFMiD of 58, highlighting their consistent co-occurrence within the dataset. These frequent combinations underscore key associations and recurring patterns, providing valuable insights for further analysis and decision-making processes. Frequent combinations, as indicated by their higher RFMiD2 values and frequent appearance within the dataset, serve as key indicators of common patterns or categories. Among these, combinations such as ('ODC', 'TSLN'), ('MH', 'ODC'), and ('MH', 'TSLN') stand out prominently. These combinations, with RFMiD2 values of 4 and 3 respectively, signify recurrent associations between specific labels. For instance, the combination ('ODC', 'TSLN') appears consistently, suggesting a notable relationship or co-occurrence between entities labeled 'ODC' and 'TSLN'. Similarly, ('MH', 'ODC') and ('MH', 'TSLN') highlight recurring connections involving the label 'MH' alongside 'ODC' and 'TSLN' respectively. Interestingly, ('MYA' [29], 'ODC') emerges as one of the most prevalent combinations with an RFMiD2 value of 9, indicating a particularly strong association between the labels 'MYA' and 'ODC'. This combination suggests a significant pattern within the dataset, possibly representing a distinct category or relationship of high importance. Also in addition, ('MH', 'MYA') also appears frequently with an RFMiD2 value of 3, further underscoring the recurrent association between 'MH' and 'MYA' labels. These frequent combinations collectively provide valuable insights into prevalent patterns or relationships within the dataset, guiding further analysis and interpretation.
- Rare combinations: among the various combinations analyzed, one particularly stands out as exceedingly rare in RFMiD dataset is the pairing of 'EDN' and 'HR' [30], which does not appear to occur at all or has a frequency count of 0. Similar other label combination namely ('MYA', 'TSLN'), ('CWS', 'HR'), ('CWS', 'EDN'), ('CME', 'HR'), ('CWS', 'EDN', 'HR') also have their combinations value at 0 thereby indicating that there are no combinations of these disease present in the current patient diagnosis. A notable point in the analysis is that there are no rare combinations that was identified in RFMiD2 dataset thereby indicating that no lable combinations had a value that accounted to 0 in RFMiD2. Also another notable analysis that was found in this study was the label combination ('EDN', 'HR'), ('MYA', 'TSLN'), and ('CWS', 'HR'), exhibit RFMiD2 values of 34, 18, and 18, respectively. This differs from the RFMiD dataset where these label combinations had their value to be 0. This is an interesting scope of study where the research can be enhanced as to find out what are the other factors that are accounted leading to this behavioural pattern where in one dataset the label combinations are found to be rare and in another found

in mid-range value.

- Isolated conditions: an isolated combination observed within the RFMiD dataset is 'ERM' and 'ODE', which appears only once, denoted by a frequency count of 1. Similiarly ('ODE', 'TSLN') also have an value of 1 indicating an isolated behaviour of the label combination. ('BRVO', 'DR', 'TSLN'), ('EDN', 'MH', 'TSLN'), ('DN' [31], 'MH', 'ODC'), ('AH', 'LS'), ('LS', 'ST'), ('CRS', 'MYA', 'ODC'), ('BRVO', 'LS', 'ST'), ('ARMD', 'EDN', 'MYA'), ('DR', 'ST') are some of the label combinations in RFMiD dataset that have an count of 1. Similarly isolated combinations, in RFMiD2 dataset characterized by their low values and infrequent occurrence within the dataset, often denote unique or niche categories that stand apart from more common patterns. Examples such as ('ERM', 'ODP'), ('EDN', 'MYA'), and ('CRS', 'ODP') exhibit values of 1, signifying their rarity and isolated nature. These combinations may represent specialized associations or uncommon co-occurrences within the dataset. The isolated combinations can be extended in the future to find out what are the factors causing these diseases to occur in less number thereby helping to take precautionary patient care.
- Complex combinations: complex label combinations indicate the presence of more 3 or more disease prevalent in the patient. While anlaysis the RFMiD and RFMiD2 dataset several label combinations within the dataset involve multiple retinal conditions, such as (ARMD, EDN, MYA), (ARMD, EDN, TSLN), (ARMD, MH, ODC), (ARMD, AH, ODC, TSLN), (AH, ARMD, ODC, TSLN), (ARMD, BRVO, MYA) and (AION, EDN, ODE). The disease combinations of 3 and gives an interesting future research scope to find out what are the factors attributing this phenomenon. A further study on a larger dataset will help preventive and timely diagnosis for the patients. The Figures 7 and 8 respectively gives an overview about the top 10 occurrences of label combination for both the datasets.

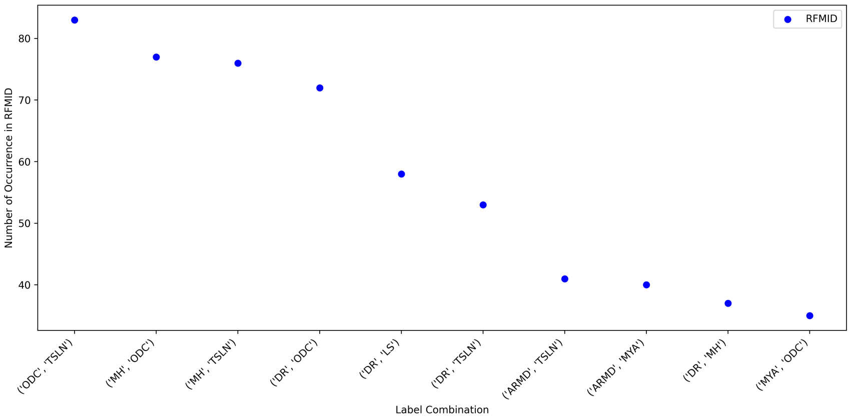

Figure 7. Scatter plot depicting the top 10 label combination occurrences of RFMiD

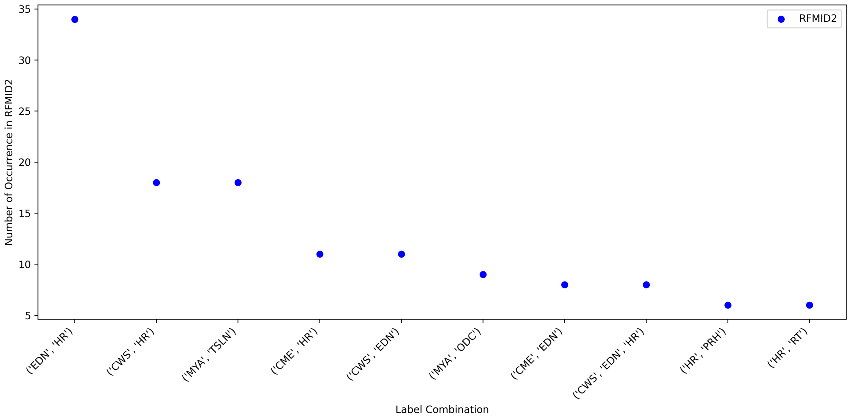

Figure 8. Scatter plot depicting the top 10 label combination occurrences of RFMiD2

3.4. Common label combination analysis

Recognizing shared label combinations across both datasets holds paramount significance, indicating common patterns and behaviors among labels across datasets. These recurrent combinations provide valuable insights for deeper investigations, facilitating a comprehensive understanding of their consistent occurrences. Furthermore, conducting comprehensive combination studies enables seamless cross-dataset analyses and contributes to the smooth integration and alignment of datasets, thereby enhancing their overall utility. The analysis underscores the prevalence of common label combinations, illuminating frequently occurring pairs. Notably, ODC and TSLN emerge prominently in both the RFMiD and RFMiD2 datasets, with respective frequencies of 83 and 4. Furthermore, the combination (MYA, ODC) is evident across the two datasets, occurring 35 and 9 times, respectively. Moreover, exploring the relationships between these shared label combinations can uncover nuanced insights into the underlying dynamics of the datasets. For instance, examining the co-occurrences of labels such as ('MH', 'ODC') and ('MH', 'TSLN') may reveal interrelated patterns in the dataset, potentially indicating specific associations or dependencies between these categories. Additionally, identifying recurring label combinations can inform targeted strategies for data integration and analysis. By focusing on these commonly occurring pairs or triples, analysts can prioritize efforts to harmonize datasets and align their structures effectively, thereby facilitating more robust comparative analyses and yielding deeper insights into the shared characteristics or trends across datasets. In essence, leveraging the knowledge of shared label combinations not only enriches our understanding of individual datasets but also empowers us to extract meaningful insights that transcend individual domains, paving the way for more holistic and impactful data-driven decision-making processes. The Figure 9 representation throws a brief overview about the label combinations common to both the datasets.

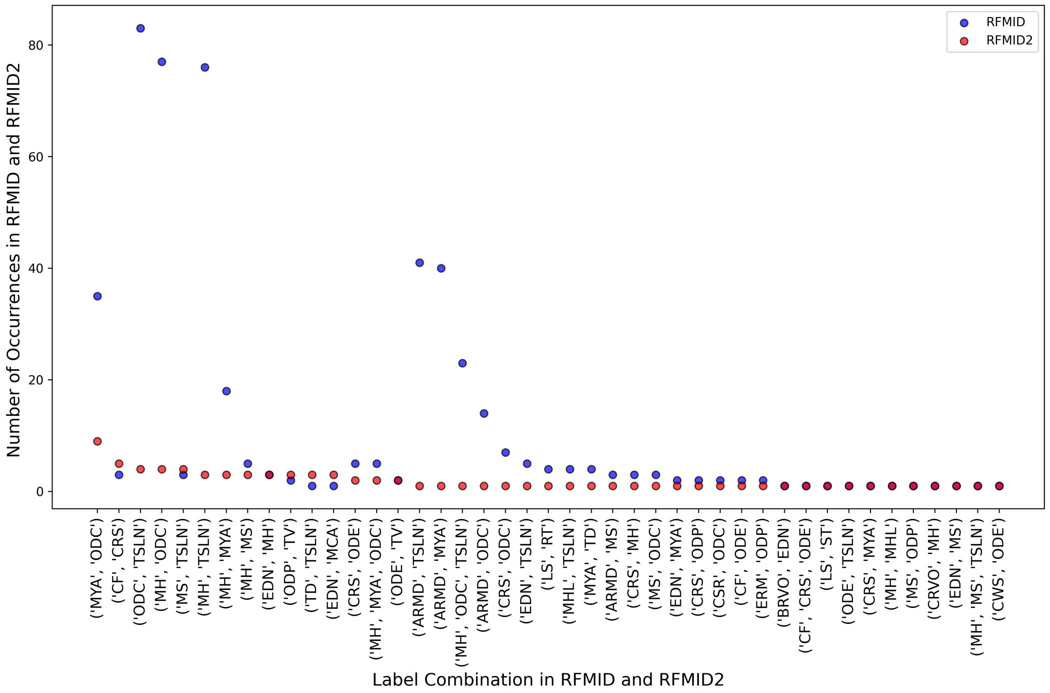

Figure 9. Scatter plot depicting label combinations common to both datasets

4. CONCLUSION

This research focused on the analysis of two significant datasets comprising retinal disease data, referred to as ”RFMiD” and ”RFMiD2.” The goal was to gain insights into the characteristics of retinal diseases, including their prevalence, associations, and common patterns. The analysis yielded several noteworthy findings and implications within the domain of retinal disease research. In summary, this research contributes to the understanding of retinal diseases by providing insights into disease prevalence, associations, and common patterns. The identification of shared disease profiles between RFMiD and RFMiD2 highlights opportunities for cross-dataset validation and the development of more effective treatment and management strategies for retinal diseases. Future research in this domain may involve clinical validation of findings and the integration of additional medical data sources to enhance diagnostic accuracy and patient care. These insights hold significant potential for improving the diagnosis and treatment of retinal diseases, ultimately benefiting patients and healthcare providers.

REFERENCES

[1] R. H. Guymer and T. G. Campbell, "Age-related macular degeneration," *The Lancet*, vol. 401, no. 10386, pp. 1459–1472, Apr. 2023, doi: 10.1016/S0140-6736(22)02609-5.

[2] T. -E. Tan and T. Y. Wong, "Diabetic retinopathy: looking forward to 2030," *Frontiers in Endocrinology*, vol. 13, Jan. 2023, doi: 10.3389/fendo.2022.1077669.

[3] A. K. Schuster, C. Erb, E. M. Hoffmann, T. Dietlein, and N. Pfeiffer, "The diagnosis and treatment of glaucoma," *Deutsches Ärzteblatt international*, vol. 117, no. 13, Mar. 2020, doi: 10.3238/arztebl.2020.0225.

[4] X. Ren *et al.*, "Artificial intelligence to distinguish retinal vein occlusion patients using color fundus photographs," *Eye*, vol. 37, no. 10, pp. 2026–2032, Jul. 2023, doi: 10.1038/s41433-022-02239-4.

[5] S. Panchal *et al.*, "Retinal fundus multi-disease image dataset (RFMiD) 2.0: a dataset of frequently and rarely identified diseases," *Data*, vol. 8, no. 2, Jan. 2023, doi: 10.3390/data8020029.

[6] A. Bhati, N. Gour, P. Khanna, and A. Ojha, "Discriminative kernel convolution network for multi-label ophthalmic disease detection on imbalanced fundus image dataset," *Computers in Biology and Medicine*, vol. 153, Feb. 2023, doi: 10.1016/j.compbiomed.2022.106519.

[7] S. Pachade *et al.*, "Retinal fundus multi-disease image dataset (RFMiD): a dataset for multi-disease detection research," *Data*, vol. 6, no. 2, Feb. 2021, doi: 10.3390/data6020014.

[8] M. Nawaz *et al.*, "Unraveling the complexity of optical coherence tomography image segmentation using machine and deep learning techniques: a review," *Computerized Medical Imaging and Graphics*, vol. 108, Sep. 2023, doi: 10.1016/j.compmedimag.2023.102269.

[9] M. D. Varela et al., "Artificial intelligence in retinal disease: clinical application, challenges, and future directions," *Graefe's Archive for Clinical and Experimental Ophthalmology*, vol. 261, no. 11, pp. 3283–3297, Nov. 2023, doi: 10.1007/s00417-023-06052-x.

[10] R. Thanki, "A deep neural network and machine learning approach for retinal fundus image classification," *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2023.100140.

[11] C. D. Vente et al., "AIROGS: artificial intelligence for robust glaucoma screening challenge," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 542–557, Jan. 2024, doi: 10.1109/TMI.2023.3313786.

[12] Z. Cai, L. Lin, H. He, and X. Tang, "Uni4Eye: unified 2D and 3D self-supervised pre-training via masked image modeling transformer for ophthalmic image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022, pp. 88–98, doi: 10.1007/978-3-031-16452-1_9.

[13] A. C. Oganov *et al.*, "Artificial intelligence in retinal image analysis: development, advances, and challenges," *Survey of Ophthalmology*, vol. 68, no. 5, pp. 905–919, Sep. 2023, doi: 10.1016/j.survophthal.2023.04.001.

[14] F. Antaki, R. G. Coussa, G. Kahwati, K. Hammamji, M. Sebag, and R. Duval, "Accuracy of automated machine learning in classifying retinal pathologies from ultra-widefield pseudocolour fundus images," *British Journal of Ophthalmology*, vol. 107, no. 1, pp. 90–95, 2023, doi: 10.1136/bjo-2022-323002.

[15] C. Danese *et al.*, "The impact of artificial intelligence on retinal disease management: vision academy retinal expert consensus," *Current Opinion in Ophthalmology*, vol. 34, no. 5, pp. 396–402, Sep. 2023, doi: 10.1097/ICU.0000000000000980.

[16] L. F. Nakayama *et al.*, "A Brazilian multilabel ophthalmological dataset (BRSET)," *PhysioNet*, 2023, doi: 10.13026/xcxw-8198.

[17] Q. Zhou, H. Zou, and Z. Wang, "Long-tailed multi-label retinal diseases recognition via relational learning and knowledge distillation," *in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, 2022, pp. 709–718, doi: 10.1007/978-3-031-16434-7_68.

[18] O. Akinniyi, M. M. Rahman, H. S. Sandhu, A. El-Baz, and F. Khalifa, "Multi-stage classification of retinal OCT using multi-scale ensemble deep architecture," *Bioengineering*, vol. 10, no. 7, Jul. 2023, doi: 10.3390/bioengineering10070823.

[19] A. Choudhary, S. Ahlawat, S. Urooj, N. Pathak, A. L. -Ekuakille, and N. Sharma, "A deep learning-based framework for retinal disease classification," *Healthcare*, vol. 11, no. 2, Jan. 2023, doi: 10.3390/healthcare11020212.

[20] O. Mosenzon, A. Y. Y. Cheng, A. A. Rabinstein, and S. Sacco, "Diabetes and stroke: what are the connections?," *Journal of Stroke*, vol. 25, no. 1, pp. 26–38, Jan. 2023, doi: 10.5853/jos.2022.02306.

[21] J. B. Park and A. Avolio, "Arteriosclerosis and atherosclerosis assessment in clinical practice: methods and significance," *Pulse*, vol. 11, no. 1, pp. 1–8, 2023, doi: 10.1159/000530616.

[22] H. Lin *et al.*, "Age and the relative importance of liver-related deaths in nonalcoholic fatty liver disease," *Hepatology*, vol. 77, no. 2, pp. 573–584, Feb. 2023, doi: 10.1002/hep.32633.

[23] P. G. Pavani, B. Biswal, and T. K. Gandhi, "Simultaneous multiclass retinal lesion segmentation using fully automated RILBP-YNet in diabetic retinopathy," *Biomedical Signal Processing and Control*, vol. 86, Sep. 2023, doi: 10.1016/j.bspc.2023.105205.

[24] V. M. Kanukollu and S. S. Ahmad, *Retinal hemorrhage*, Treasure Island, Florida: StatPearls Publishing, 2023.

[25] M. Jain *et al.*, "Post-vitrectomy secondary macular holes: risk factors, clinical features, and multivariate analysis of outcome predictors," *Indian Journal of Ophthalmology*, vol. 71, no. 5, pp. 2053–2060, May 2023, doi: 10.4103/ijo.IJO_1749_22.

[26] K. Nastiti, P. R. A. Sangging, and R. Himayani, "Optic disc cupping," *Medical Profession Journal of Lampung*, vol. 13, no. 4.1, pp. 207–208, 2023.

[27] S. Napoli, M. Zanardelli, A. M. D'Erme, and T. M. Lotti, "Sclerotherapy," *in European Handbook of Dermatological Treatments*, Cham: Springer International Publishing, 2023, pp. 1449–1454, doi: 10.1007/978-3-031-15130-9_129.

[28] M. Hein, A. Mehnert, K. B. Freund, D.-Y. Yu, and C. Balaratnasingam, "Variability in capillary perfusion is increased in regions of

retinal ischemia due to branch retinal vein occlusion," *Investigative Opthalmology & Visual Science*, vol. 64, no. 13, Oct. 2023, doi: 10.1167/iovs.64.13.30.

[29] J. B. Jonas, R. A. Jonas, M. M. Bikbov, Y. X. Wang, and S. Panda-Jonas, "Myopia: histology, clinical features, and potential implications for the etiology of axial elongation," *Progress in Retinal and Eye Research*, vol. 96, Sep. 2023, doi: 10.1016/j.preteyeres.2022.101156.

[30] P. Fu *et al.*, "Efficacy and safety of pan retinal photocoagulation combined with intravitreal anti-VEGF agents for high-risk proliferative diabetic retinopathy: a systematic review and meta-analysis," *Medicine*, vol. 102, no. 39, Sep. 2023, doi: 10.1097/MD.0000000000034856.

[31] C. Arunavinodhini and S. Sabena, "DREXUNET: a novel deep CNN algorithm for drusen and exudates classification on fundus images," *Research Square*, pp. 1-17, 2023, doi: 10.21203/rs.3.rs-3383744/v1.

BIOGRAPHIES OF AUTHORS

Sridhevi Sundararajan is actively engaged in her postgraduate studies, pursuing a Master of Technology in Engineering Design at Symbiosis Institute of Technology in Pune, India. She has around 15 years of experience in IT domain with C++, unix being her main forte. Her area of interest is research in medical domain. She likes to spend her time reading books related to medical field and exploring unknown. She can be contacted at email: sridevisundarajan@gmail.com.

Harikrishnan Ramachandran is currently working as associate professor in the Department of Electronics and Telecommunication Engineering, Symbiosis Institute of Technology, Pune Campus, Symbiosis International Deemed University, Pune, India. His main research interest includes smart grid, optimization, internet of things, artificial intelligence, and data analytics. He is an IEEE senior member, fellow life member of Institution of Engineers India, fellow life member of IETE India, life member of Indian Society for Technical Education and life member of Computer Society of India. He can be contacted at email: dr.rhareish@gmail.com.

Harshita Gupta is currently in the final year of her bachelor's degree in Electronics and Telecommunication Engineering from Symbiosis Institute of Technology, Pune, Symbiosis International Deemed University. Additionally. She is pursuing a minor in Data Science from the Department of Computer Science at the same university. With a keen interest in the dynamic fields of artificial intelligence, machine learning, deep learning, and computer vision. She is poised to contribute to the future of technology. She can be contacted at email: harshitag2810@gmail.com.