CONTENT-BASED LECTURE VIDEO RETRIEVAL

YİĞİT ŞAHİN

FEBRUARY 2023

ÇANKAYA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

M.Sc. Thesis in

COMPUTER ENGINEERING

CONTENT-BASED VIDEO RETRIEVAL

YİĞİT ŞAHİN

FEBRUARY 2023

ÖZET

DERS VİDEOLARINA İÇERİK TABANLI ERİŞİM

ŞAHİN, YİĞİT
Bilgisayar Mühendisliği Yüksek Lisans

Danışman: Prof. Dr. Hasan Oğul
Şubat 2023, 76 Sayfa

Günümüzde e-öğrenme veya çevrimiçi öğrenme olarak sıklıkla karşımıza çıkan uzaktan eğitim, eğitim-öğretim sırasında eğitmen ile öğrencinin fiziksel olarak yan yana olmadığı ve öğrenci-eğitmen iletişimini kolaylaştırmak için çeşitli teknolojilerin kullanıldığı yeni nesil bir eğitim yaklaşımıdır. Bu yaklaşım koronavirüs pandemisi (COVID-19) ile dünya çapındaki ağda (www) özellikle ders-eğitim içerikli videolar ile daha yaygın ve kullanılabilir hale gelmiştir. Ancak internet ortamında bulunan video sayılarındaki yüksek artış hızı, belirli bir içeriğe sahip videoya ulaşmak isteyen kullanıcıların video içeriklerine erişimini oldukça zorlaştırmıştır. Bahsedilen bu zorluklara bir öneri geliştirmek bağlamında bu araştırmada kullanıcıların belirli eğitim içerikleri ile ilgili videolara erişimini amaçlayan içerik tabanlı erişim yöntemi ele alınmıştır.

Kullanıcıların aradıkları video içeriklerine daha kolay ulaşması için videoların doğru sınıflandırılması gerekmektedir. Teknik açıdan sınıflandırılmanın yapılabilmesi için ise öncelikle videoların metin bilgilerine ulaşılmalıdır. Bu çerçevede araştırmada videoların metin bilgilerini çıkarmak için optik karakter tanıma (OCR) ve otomatik konuşma tanıma (ASR) isimli iki farklı indeksleme yöntemi kullanılmıştır. Bu iki yöntem ve bu iki yöntemin birlikte kullanıldığı bir analiz, bu tezde belirli bir veri kümesi üzerinden ele alınmıştır. Veri kümesi olarak ise 110 videolu bir eğitim koleksiyonu kullanılmıştır. Bu kapsamda aynı veriyi kullanarak OCR ile analiz yapmış bir tez referans alınarak, bu kez de ASR yöntemi ile aynı metrik

analizler yapılmıştır. Son olarak ise, hem OCR hem de ASR yöntemi kullanılarak çeşitli metrik değerler hesaplanmıştır. Verilerin sınıflandırma analizi için 3 farklı geleneksel makine öğrenme yöntemi kullanılmıştır. Kullanılan geleneksel makine öğrenme yöntemleri Support Vector Machine (SVM), Naïve Bayes ve Random Forest yöntemleridir. Bu doğrultuda farklı makine öğrenme yöntemleri ve farklı indeksleme yöntemlerinin aynı veri kümesi üzerinde metrik analizleri karşılaştırılmıştır. Yapılan analizler sonucunda ders videolarının içerik tabanlı erişimde kullanılabilmesi için günümüzde mümkün olan geleneksel makine öğrenme yöntemleri ile indeksleme yöntemlerinin güçlü ve zayıf yönlerinin açıklaması yapılmıştır. Bununla birlikte aynı konuda yapılacak gelecek çalışmalar için yöntemin geliştirilebilecek yönleri vurgulanmış ve bu konudaki öneriler sunulmuştur. Bu tez, yapılan karşılaştırmalı araştırmanın hem eğitim hem de yazılım sektörünü nasıl etkileyebileceği tartışması ile noktalanmaktadır.

Anahtar Kelimeler: Ders Videolarının İçerik Tabanlı Erişimi, Support Vector Machine, Naïve Bayes, Random Forest, Optik karakter tanıma, Otomatik Konuşma Tanıma

ABSTRACT

CONTENT BASED LECTURE VIDEO RETRIEVAL

ŞAHİN, YİĞİT

M.Sc. in Computer Engineering

Supervisor: Assoc. Prof. Dr. Hasan Oğul

February 2023, 76 Pages

Distance education, which is frequently encountered as e-learning or online learning today, is a new generation education approach in which the instructor and the student are not physically in the same place during education. Various technologies are used to facilitate online student-instructor communication. This approach has become more common and available on the world wide web (www) with the coronavirus pandemic (COVID-19), especially regarding lecture videos. However, the high rate of increase in the number of lecture videos on the internet has made it very difficult for users who want to access a specific video with a certain content. In the context of developing a proposal for these challenges, this research deals with the content-based search method that aims to provide users with access to videos related to certain educational content.

The videos need to be accurately categorized in order to make it easier for users to find the specific video content. From the technical point of view, in order to be able to make such categorization, the textual information of the videos should be first retrieved. In this context, two different indexing methods called optical character recognition (OCR) and automatic speech recognition (ASR) are adopted to extract the textual information of the videos in the research. These two methods and an analysis in which these two methods are used together, are handled in this thesis over a specific data set. A collection of 110 lecture videos is used as the dataset. Within this scope, by referring to a thesis that made analysis with OCR method using the same dataset,

this time the same metric analyzes are made with the ASR method. Finally, various metric values are calculated using both OCR and ASR methods. Three different traditional machine learning methods are used for classification analysis of the dataset. The traditional machine learning methods used are Support Vector Machine (SVM), Naïve Bayes and Random Forest methods. Accordingly, metric analysis of different machine learning methods and different indexing methods on the same dataset are compared. As a result of the analysis, the strengths and weaknesses of traditional machine learning methods and indexing methods are explained in order to use the lecture videos in content-based search. In addition, the aspects of the method that can be improved for future studies on the same subject are emphasized and suggestions on this subject are presented. This thesis concludes with a discussion of how the comparative research can affect both the education and software industries.

Keywords: Content-Based Lecture Video Retrieval, Support Vector Machine, Naïve Bayes, Random Forest, Optical Character Recognition, Automatic Speech Recognition

ACKNOWLEDGMENT

I sincerely thank the following individuals for their contributions to the realization of this study:

I would like to express my deepest appreciation to Prof. Dr. Hasan Oğul, my thesis advisor, for always helping and guiding me in the completion of the study and overcoming the difficulties encountered. Special thanks to Veysel Sercan Ağzıyağlı for all the guidance and inspiration. In addition, I'm extremely grateful to Sema Ceran for her invaluable patience and feedback.

Furthermore, I could not have completed this journey without the unconditional support from my family. Lastly, I would like to express my gratitude for Defne Işıklı whose support has been remarkable throughout this thesis.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ABBREVIATIONS

API	: Application Programming Interface
ASR	: Automatic Speech Recognition
BIC	: Bayesian Information Criterion
Bi-LSTM	: Bidirectional Long Short-Term Memory
CNN	: Convolutional Neural Network
COVID-19	: Coronavirus Pandemic in 2019
HOG	: Histogram of Oriented Gradients
LMS	: Learning Management System
MOOC	: Massive Open Online Courses
NLP	: Natural Language Processing
OCR	: Optical Character Recognition
OLAT	: Online Learning and Training
SCORM	: Sharable Content Object Reference Model
SVM	: Support Vector machine
US	: United States
VAD	: Voice Activity Detection
WWW	: World Wide Web

CHAPTER I
INTRODUCTION

Over the years, technology has revolutionized our world and daily lives. Especially the development of technology has brought radical changes to the education system. Specifically, with the development of systems such as tablets, portable computers, fiber internet infrastructures, and online education platforms, online education has become widely reachable and possible. Therefore, online education is gradually starting to replace face-to-face conduct in the educational system and will be more dominant as an educational method of conduct in the upcoming years.

Elliott Masie, a United States (US) learning expert, first used the word "e-learning" in his November 1999 keynote address at the “TechLearn Conference”. He mentioned that "e-learning is the use of network technology to design, deliver, select, administer, and extend learning." [1]. The first open-source learning management system (LMS), online learning and training (OLAT) was released in 2000 and revolutionized e-learning. The first iteration of the sharable content object reference model (SCORM) standard, which enabled users to package and share content inside the LMS, was also made available the same year [2]. Mobile gadgets, such as smartphones and tablets, started to proliferate in the early 2000s. People utilized mobile phones for more than just making phone calls; they also watched videos, read books, and played games [3]. This sparked a race among cellular firms to enhance mobile connectivity and interaction, which is still strongly present today. Accessing e-learning from mobile devices has already become a common choice for corporations, institutions, and organizations.

Even though different nations have varied rates of COVID-19 infection, the pandemic has caused school closures in 186 different nations that affect more than 1.2 billion children worldwide. With global education technology investments exceeding US$18.66 billion in 2019 and the whole market for online education estimated to reach $350 Billion by 2025, there was already substantial growth and adoption in

education technology before COVID-19 [4]. Since COVID-19, there has been a noticeable increase in the utilization of language apps, virtual tutoring, video conferencing tools, and online learning software.

The epidemic has proved that online learning is a viable and sustainable paradigm. It maintains access to education during a public health emergency, natural disaster, or other situations in which students and teachers cannot come together in the same physical space. Distance learning can serve a broader variety of pupils, regardless of geographical status, by incorporating voices from throughout the country and beyond.

The increased tendency of utilizing online learning methods throughout educational systems resulted in increasing the number of educational videos. The process of searching and finding the corresponding video content has become more complicated. In order to search online lectures from repositories, many different methods have been developed, but videos have rich defining content inside. Therefore, it can be hard to reach the explanatory audio-visual content only from the name of the lecture, the description, the teacher, and the curriculum. The reason why this research is needed is precisely because this search method has become difficult. Consequently, a video content base search method is used in this thesis to increase the accuracy of searching and retrieving videos from databases.

Based on the phenomenon mentioned above, for this thesis, 110 videos are compiled into a dataset. This dataset contains a total of 110 videos, as well as distinct types and hierarchical layers of video material. Examining what kind of datasets are more successful when employing distinct contents at different hierarchical levels.

Little research has been conducted on video content-based search. Some of the research is going to be compared in the section on the literature review. This thesis is distinctive since it uses both ASR and OCR on the same dataset to analyze their compatibility with traditional classification algorithms. This thesis will throw light upon which indexing method with traditional classification algorithm can provide more accurate search results to the companies that produce educational videos dealing with streaming.

The primary aim of this thesis is to examine, analyze and compare ASR, OCR, and both ASR and OCR algorithms utilizing various metrics on the dataset. In accordance with Ağzıyağlı thesis, who has used the exact 110 videos, he had already demonstrated the OCR method's results [5]. The same dataset is then utilized to

generate the ASR method dataset. As a result, results are achieved using both ASR and OCR methods. These three research all uses the same data set.

This thesis uses the methods shown below and compared these methods under various metrics.

- ASR
- OCR
- ASR and OCR

The word frequency vectors form the foundation of the method that underpins the outputs of both the ASR and OCR. Textual information obtained from ASR and OCR is calculated with normalization and non-normalization in order to compare them. Using three different approaches to machine learning, these vectors are sorted into their respective categories. In this particular instance, the machine learning strategies of Naïve Bayes [6], Random Forest [7], and SVM [8] are utilized. All three of the traditional approaches to machine learning have been trained and assessed through the utilization of frequency vectors.

The proposed methods are evaluated using a wide range of different volumes of lecture videos, each of which features a distinct level of semantic depth and hierarchical structure. After putting preprocessing algorithms through their paces, three different categorization approaches have been compiled. In total, there are 110 that are employed. These 110 videos have been arranged in a hierarchical structure into various levels of significance.

1.1. MOTIVATION

With the digitalization of educational tools, educational institutions are adopting online learning methods and gradually breaking the conventional face-to-face teaching tendencies. Therefore, universities and such learning institutions are becoming more than a physically reachable entity. University education is evolving into being in a non-linear form; where the participants can watch/view the course content and access the source material at any desired time. Subsequently, the online education surpasses the limitations of physicality by becoming a more rational, modern, and innovative system. Distance learning, which is completely independent of time and place, without the requirement for the student and lecturer to meet in the same physical environment, through the existing computer technologies, live, video, audio, and interactive lessons, are taught in a completely virtual environment.

The increasing technology and spread of the COVID-19 pandemic have accelerated the development of distance learning. Since the introduction of online education, there has been a rise in the number of video lectures that are stored in online archives. At this juncture, the question of how to obtain the video recordings of these lectures with the highest possible degree of specificity occurs. It arises a question which is addressing the curiosity of both educational institutions and individuals.

In order to investigate such curiosity, lecture videos were selected as the main subject of focus of the research of this thesis. Mainly, the thesis aims to classify the lecture videos based on to the content since, classifying the lecture videos cannot be always enough with the lecture name, name of the teacher, or lecture description. These videos can have richer content than the lecture name, name of the teacher, or lecture description. In addition, the thesis seeks answers to which indexing method has a higher accuracy rate for lecture videos in order to help the streaming sector.

This thesis contributes to the body of knowledge by utilizing many classification methods on ASR-obtained data from distinct a distinct dataset and determining which indexing approach (ASR or OCR) provides the most accurate results for lecture videos.

1.2 RESEARCH QUESTIONS AND OBJECTIVES

There are several characteristics that distinguish lecture videos from other sorts of videos; typically, they include text content, video frames, and audio tracks [10]. Mainly, this thesis focuses on the speech within the lecture videos and how to extract the speech from the lecture videos as the textual information with ASR method.

Due to the scarcity of research on lecture videos in the academic literature, it is also essential to evaluate which indexing approach is the most effective. In this thesis, the findings generated by the OCR method are evaluated [5] and the ASR method is applied to the same dataset using the same results. Consequently, OCR, ASR, and both (OCR and ASR combined) approaches are implemented on the same dataset, and the outcomes are compared and evaluated.

This thesis is motivated by two main research questions:

1. Which indexing method works more efficiently on lecture videos?
2. How can lecture videos be classified using the ASR method?

These questions require further examinations through these queries:

1. How can textual information be extracted from lecture videos?

2. How can the stop word list be determined?
3. Which tools should be used in this project?
4. How to compare OCR and ASR methods in some metrics?
5. How to create the dataset?
6. How to create and use the Weka dataset?

1.3 ORGANIZATION OF THESIS

The organization of this thesis is as follows: The first chapter discusses the project's motivation and scope briefly. The second chapter examines the relevant literature and previous studies on this topic. The dataset that has been utilized for the analysis is outlined in the third chapter. The methods of indexing that have been applied to the project are the topic of discussion in the fourth chapter. This chapter also provides detailed information regarding the matter. In the fifth chapter, typical classification techniques have been used to classify text outputs are described. In the sixth chapter, the methodology and general structure of the thesis are described and explained. The results of the experiments that have been conducted for the thesis are illustrated in the seventh chapter. The outcomes of the classification algorithms used are discussed in Chapter 8. In Chapter 9, results and potential future research are discussed.

CHAPTER II
LITERATURE REVIEW

Digital audiovisual records are being used extensively in today's educational system because of their accessibility online, independence from location, and portability. It is not easy to find relevant videos on the internet that pertain to a particular subject. E-learning material needs to be developed quickly so that content-based lecture videos may be found more quickly. For this reason, research on content-based lecture videos is increasing rapidly.

Ağzıyağlı [5] has published an analysis paper on the search process for lecture videos, even though there are not many studies in the literature that compares the performance of different algorithms on textual data for various lecture videos, as in this thesis. The research of Ağzıyağlı [5] has highlighted the advantages of adopting content-based search. Moreover, this investigation has shed considerable information on this thesis; since it is one of the few types of research that applies to lecture videos. To compare the performance of OCR with ASR methods, the dataset of Ağzıyaglı's [5] work has been utilized in this thesis.

There is a wealth of content in videos, including images, text, and speech, which cannot be adequately represented by metadata alone. There is an increasing demand for the development of technologies that automatically annotate lecture videos in order to assist in conducting more precise research. In the dissertation by Ağzıyağlı [5], a textual technique and numerous algorithms that can be applied with this textual approach are proposed to classify lecture videos based on their content in a more accurate manner.

A dataset consisting of 110 videos is constructed in order to support the claims made in Ağzıyağlı's [5] thesis. This dataset is comprised of a total of 110 videos, each of which is organized into a different category and level of hierarchical content. An investigation on which types of datasets have more accurate results when applying separate contents at various hierarchical levels has been conducted.

Textual evidence forms the basis of the methods based on Ağzıyağlı's thesis [5]. In order to extract textual information from lecture videos, for Ağzıyağlı's [5] thesis, the technology known as OCR is utilized. OCR technology converts the spoken content of videos into a textual format. Following OCR, two separate preprocessing methods are used for the data. The utilization of word frequency vectors is important to one of the two unique preprocessing procedures. Using three different approaches to machine learning, the vectors were sorted into their respective categories.

The generation of vectors for word sorting is the focus of the second of two independent preprocessing approaches. In this section, the OCR-extracted words are vectorized one after the other. One of the approaches utilized in deep learning is called the "Bidirectional Long Short-Term Memory" (Bi-LSTM) algorithm. This algorithm receives these vectors in a progressive manner.

For the purpose of evaluating the proposed methods, a wide range of volumes of videos featuring a variety of semantic levels and hierarchical structures are used. After putting preprocessing algorithms through their paces, four different categorization methodologies have been examined. In total, 110 videos are employed. These 110 videos have been arranged in a hierarchical structure into various levels of significance. The efficiency of the system is based using several different measures of classification precision.

In Ağzıyağlı's dissertation research [5], there are three distinct groups of datasets. The extracted data are preprocessed using three distinct threshold settings for the similarity ratio. The application of three distinct similarity ratio criteria has produced a total of nine distinct datasets. The accuracy rates of algorithms utilized in datasets vary based on the datasets.

Level 1 and level 2 datasets are the least balanced of the nine obtained datasets. In terms of accuracy, sensitivity, precision, and F1 score ratios, the Naïve Bayes and Support Vector Machine algorithms have performed better on the level 1 dataset with a similarity criterion of 50%.

Based on the precision and F1 score ratios of the Random Forest technique, level 1 datasets are incomputable. The level 1 dataset has four classes, with 76% of clips belonging to a single class. In the level 1 dataset, the Random Forest method is unable of producing a classification whose precision and F1 score can be assessed.

When evaluating the long short-term memory approach in terms of accuracy, sensitivity, precision, and F1 score rates, datasets with a similarity rate threshold of

50% have performed the worst. This is due to the textual data lost when the threshold value is decreased.

According to Ağzıyağlı's thesis [5], in terms of accuracy, sensitivity, precision, and F1 score ratios, the Naïve Bayes method is the most successful classical machine learning method.

Taking the arithmetic mean of the results has obtained in terms of accuracy, sensitivity, precision, and F1 score ratios, the long short term memory approach has proved to be the most effective classification technique among the four machine learning methods evaluated. In comparisons based on the arithmetic mean of all measurement techniques, the long short term memory technique yielded superior results.

In their dissertation, Chand and Oğul [9] have examined and analyzed lecture videos. In addition, they have emphasized in their thesis that the educational method has evolved and is currently attempting to accommodate new tendencies.

In Chand and Oğul's thesis [9], a lecture video segmentation model based solely on the speech content of instructors is given. This project [9] aims to extract textual and acoustic elements from the audio taken from lecture videos in order to segment the lecture video. Among the most important reasons for accomplishing so is that, unlike other sources that may or may not be available and usable, lecture videos always include an audio track. To achieve this objective, a variety of open-source tools and algorithms, such as audio extractor, voice activity detection (VAD), ASR, acoustic feature extractor, and segmentation algorithms, are utilized because they are readily accessible and free of charge, and there is always a great deal of resources available when employing them.

Chand and Oğul's thesis [9] is successful in developing a system for segmenting lecture videos based entirely on the content of the speaker's speech. Initially, the research is conducted on the procedures that must be completed before creating a lecture video segmentation pipeline. The first prototype has to be developed for the research to be considered successful, and this pipeline can help. After that, a number of different open-source tools and algorithms are used to find the most effective approach for separating the audio content of the lecture video. Lessons recorded from massive open online courses (MOOC) sites are used to generate a dataset and analyze the findings of the proposed prototype. In addition, the findings are compared with various conditions to evaluate the efficiency of the system. This

comparison has shown that the model has outperformed other comparable systems on all three measures: sensitivity, recall, and F-score. Considering these aspects, it is argued that the method proposed in this thesis can effectively segment lecture videos according to their speaking content.

In the thesis of Yang and Meinel [10] a highly fruitful investigation has been conducted. Yang and Meinel [10] claims that the number of educational videos available on the Internet is growing daily. Therefore, in this thesis, it is presented automated video indexing and video search in large lecture video archives. ASR and OCR technologies are also utilized to acquire textual information, similar to this thesis. Consequently, this premise is crucial for analyzing the outcomes of this thesis.

Due to the advancement of technology, it is common knowledge that distance education has become extremely ubiquitous. There are numerous videos on any topic available in web archives. The first question that arises is how to match the appropriate lecture video with the appropriate user. The majority of video retrieval and video search engines, including YouTube, Bing, and Vimeo, return results based on textual metadata such as title, genre, person, brief description, and so forth. Typically, this type of metadata must be developed by a human to assure good quality, which is a time- and cost-intensive process. Consequently, Yang and Meinel's thesis [10] is predicated mostly on the following premise, as mentioned: "Using appropriate analysis techniques, it is possible to automatically extract relevant information from lecture recordings. They can assist users in rapidly locating and comprehending course content, hence enhancing learning activity."[10].

OCR and ASR technologies are utilized to analyze the lecture videos. These technologies allow for the acquisition of vast amounts of textual data. To avoid solidity and consistency issues, the term "Frequency Inverse Document" [12] score is utilized. In addition, the "Cosine Similarity Measure" [11] approach is utilized to determine the similarity of videos.

The method of scoring data acquired from textual and aural sources as keywords are presented. In this manner, textual and acoustic data are graded according to their significance and included in the system. Once every three seconds, OCR checks the difference between frames or sections isolated from frames in the research [10]. Any modification to the reflections that occurred in less than three seconds is disregarded by the system.

SVM categorizes based on the histogram characteristics between mirrors. The study has been experimentally tested. Twenty distinct movies featuring distinct speakers are utilized in the study. In the study, the SVM classifier is trained using 2597 mirror image segments and 5224 non-mirrored image segments. Textual information in image fragments has been digitized and processed using the open-source tesseract application. The Tesseract algorithm has identified letters 92% of the time and words 85% of the time.

In the second part of Yang and Meinel's thesis [10], ASR technology is utilized. Using the "CMU Sphinx Toolkit" and the "German Speech Corpus" by Voxforge as a foundation, it is determined to develop acoustic models for a specific use case. In contrast to earlier methods that collect speech data by dictation in a calm atmosphere, it has collected hours of speech data from actual lecture recordings and generated related transcripts. Thus, the actual classroom setting can be incorporated into the training process. It has utilized the gathered text corpora from the German daily news corpus (radio programs, 1996-2000), "Wortschatz- Leipzig", and the audio transcripts of the collected speech corpora to train the language model.

In Haubold and Kender's paper [13], it is investigated methods for segmenting, displaying, and indexing presentation recordings using audio and visual data separately are examined. In a nutshell, the speaker divides the audio recording into multiple parts and then uses ASR software to pull out key phrases to add to the recording. The video part is broken up using separate visual elements as well as keyframes that are reflective of those elements. An interactive user interface makes it possible for a user to navigate around a presentation video by combining audio, video, text, and a graphical representation of keyframes. In addition, the grouping and tagging of speaker data are investigated, and preliminary conclusions are presented [13].

The investigation in Haubold and Kender's research [13] is made use of a video recording of a class presentation that lasted for seven and a half hours. In total, there are 32 presentations given. These presentations involve 176 unique students. There is a wide range of expertise present among these 176 pupils. The presentations, that take place in the classroom, are almost always given by numerous students and follow a predetermined outline. Presentations in the classroom differ significantly from videos of lectures in several significant ways. These recordings are longer than the typical classroom lectures that are shown on video, and they feature a large number of

students in the audience. The level of audio quality as well as the frequency with which certain keywords are repeated varies greatly from one presentation to another.

The segmentation of speakers is helpful for the audio and video segmentation of presentations. The strategy for identifying changes in the speaker that was proposed by the “Bayesian Information Criterion” (BIC) [14] has been put into practice. Samples are taken from the audio channel at predetermined intervals, and thirteen “Mel Frequency Cepstral Coefficient Vectors” are computed for each sample set. The BIC is calculated for each portion of this interval by employing a method with two windows. According to the conclusion obtained by the ASR output in this thesis, the presence of a net positive maximum between the BIC values indicates the presence of a loudspeaker change.

In the research conducted by Haubold and Kender [13], the students are tasked with annotating the presentations in addition to searching for both recognized and unfamiliar presentation components. The visual content and indexing methods that are utilized, are put to the test with 176 students who possessed a variety of knowledge regarding the front interface material of the movies. According to the figures, access has offered responses that were 20% faster than those provided by comparable systems.

Adcock and Cooper [15] discuss the planning and implementation that goes into developing a webcast search engine. Within the scope of Adcock and Cooper’s study [15], the videos that make up the dataset are dissected in order to unearth a wide range of slide photographs that are afterward OCR-processed and lexically processed. Researchers are able to analyze base frame differentiation strategies to extract keyframe slide photographs by combining a speaker and a presentation slide in-frame, rotating cameras, and making slides. These are some algorithms that can help enhance the identification of slides. To evaluate the functionality of the search engine and its algorithms, a prototype of a system is constructed.

Users, particularly students, want to be able to pinpoint in a lecture exactly where a certain topic is covered at a given point in time. In order to respond to these queries, a search engine that can analyze the written content of an article on a website and identify key terms is required. "TalkMiner" [15] generates a search index from the words on the presentation slides of the source video to address this issue. The algorithm examines the video in order to detect distinct slide images. Each photograph is accompanied by a code indicating when it was first viewed.

Adcock and Cooper [15] have used around 200 lecture videos and 100 non-lecture videos in their test set. With frames from the first 5 minutes of each video, an SVM classifier is trained to differentiate between lecture and non-lecture content. This uses a binary classification technique. The following properties are utilized: the overall duration of the static content, the number of static video segments, the average length of the static video segments, as well as the minimum and average entropy of the projection of the vertical edges. 1%-pixel change is used as the threshold value depending on whether the frames are fixed or not. The classification procedure is completed with a success rate of 95%. It has achieved 98% classification accuracy with a 9% rejection rate by processing the classification score as a measure of confidence and rejecting classifications with low confidence.

Consequently, TalkMiner [15] is a rich search and browsing engine intended to increase access to and utilization of course webcasts. Utilizing existing online video distribution infrastructure, the solution embeds the original webcast into an interface for easy search and navigation within the video. TalkMiner [15] accomplishes this with minimal system computation and storage.

Chivadshetti, Sadafale, and Thakare's [16] paper is another important example of a video content-based retrieval search. This study has extracted textual keywords by applying OCR technology to the video, keyframes, and ASR technology to the audio tracks of the video.

Six primary modules comprise the content-based video retrieval personalized systems proposed. First, a user uploads a video query to the content-based video retrieval personalized system or provides it as input. The system partitions the frames into video and applies a suitable frame selection method to all frames. Using ASR technology, the ASR system simultaneously processes video input and extracts the keywords. After frame segmentation and selection, it executes OCR and extracts the histogram of oriented gradients (HOG), OCR text, and Gabor Filter from the selected frames. Additionally, it extracts the color, texture, and edge detectors from the selected frames. On videos saved in a database or the cloud, the same ASR, frame segmentation, OCR, and image processing operations are performed. After preprocessing, the system looks for similarities in keywords and features between the query video metadata and all database or cloud-stored videos. The content-based video retrieval system collects the most relevant OCR text, ASR text, keywords, and features and delivers pertinent video results. It checks the user's profile history for

customizing the results then, re-ranks the results and presents them to the user. In order to make text detection, the SVM classification algorithm is utilized in this research [16].

The research's dataset consists of 15 videos which are stored in the database. To make experiments, users send out some video queries. The users search the video by using OCR, ASR, and the combination of OCR and ASR systems. Consequently, according to the outputs of the research, while using the SVM classification algorithm, the combination of the OCR and ASR techniques has the best consistent values.

CHAPTER III
DATASET

For this thesis, the same data set as Ağzıyağlı [5] are utilized. In total, 110 course videos are included in the dataset. The longest video is the 85th video at 1 hour 21 minutes and 11 seconds. The shortest video is the 50th video at 2 minutes and 22 seconds. Another point to keep in mind is that there is no speech in the 105th video. For this reason, it is not expected to have 100 percent accurate results for the ASR module. The screen ratios of the videos in the dataset are 1920x1080 or 1280x720, therefore they are all 16:9 videos. All videos are progressive and 30 fps (single pass). The collection contains videos with the ".mp4" extension. Video codecs are AVC. The dataset contains three distinct tiers. There are four classes on the first level, three on the second level, and four on the third. Detailed information about the list of lecture videos can be found in Table 1.

Table 1: List of lecture videos used for evaluation

Video ID	Type of Video	Video length (mm:ss)	Video size (KB)
Video _001	Education	46:38	396.416
Video _002	Education	18:38	40.162
Video _003	Education	08:40	18.098
Video _004	Education	14:17	221.934
Video _005	Education	07:44	29.128
Video _006	Education	16:13	40.402
Video _007	Education	07:13	27.667
Video _008	Education	13:07	26.145

Table 1 Continued

Video ID	Type of Video	Video length (mm:ss)	Video size (KB)
Video _009	Education	16:44	31.083
Video _010	Education	15:50	22.461
Video _011	Medicine	47:18	62.832
Video _012	Medicine	29:43	58.056
Video _013	Medicine	43:57	63.536
Video _014	Medicine	40:36	68.741
Video _015	Medicine	41:57	49.665
Video _016	Medicine	48:23	65.984
Video _017	Medicine	46:09	60.400
Video _018	Medicine	48:08	67.379
Video _019	Medicine	01:00:23	85.158
Video _020	Medicine	31:08	48.246
Video _021	Social	21:28	114.015
Video _022	Social	09:35	18.605
Video _023	Social	26:20	50.597
Video _024	Social	22:15	43.406
Video _025	Social	16:22	100.863
Video _026	Social	24:49	46.903
Video _027	Social	08:38	48.652
Video _028	Social	19:54	85.571
Video _029	Social	24:13	113.410
Video _030	Social	22:40	63.366

Table 1 Continued

Video ID	Type of Video	Video length (mm:ss)	Video size (KB)
Video _031	Electronic Engineering	13:57	21.755
Video _032	Electronic Engineering	12:03	17.262
Video _033	Electronic Engineering	10:17	16.883
Video _034	Electronic Engineering	11:01	17.761
Video _035	Electronic Engineering	07:19	50.551
Video _036	Electronic Engineering	07:52	61.261
Video _037	Electronic Engineering	11:39	54.614
Video _038	Electronic Engineering	06:01	50.355
Video _039	Electronic Engineering	10:51	43.005
Video _040	Electronic Engineering	11:08	49.033
Video _041	Electronic Engineering	08:26	33.994
Video _042	Electronic Engineering	10:09	43.925
Video _043	Electronic Engineering	11:07	54.801
Video _044	Electronic Engineering	14:36	71.063
Video _045	Electronic Engineering	02:41	17.730
Video _046	Electronic Engineering	03:38	42.002
Video _047	Electronic Engineering	07:02	58.185
Video _048	Electronic Engineering	02:40	27.595
Video _049	Electronic Engineering	03:15	27.718
Video _050	Electronic Engineering	02:22	17.002
Video _051	Mechanical Engineering	10:10	25.584
Video _052	Mechanical Engineering	11:42	30.237
Video _053	Mechanical Engineering	21:27	42.141
Video _054	Mechanical Engineering	07:26	14.110
Video _055	Mechanical Engineering	10:27	21.329
Video _056	Mechanical Engineering	21:58	33.286

Table 1 Continued

Video ID	Type of Video	Video length (mm:ss)	Video size (KB)
Video _057	Mechanical Engineering	19:50	42.288
Video _058	Mechanical Engineering	18:19	37.558
Video _059	Mechanical Engineering	37:16	69.159
Video _060	Mechanical Engineering	18:32	28.003
Video _061	Mechanical Engineering	35:47	62.302
Video _062	Mechanical Engineering	22:03	41.319
Video _063	Mechanical Engineering	04:42	7.569
Video _064	Mechanical Engineering	01:10:34	100.479
Video _065	Mechanical Engineering	30:30	60.042
Video _066	Mechanical Engineering	19:02	39.544
Video _067	Mechanical Engineering	09:03	17.071
Video _068	Mechanical Engineering	28:25	54.728
Video _069	Mechanical Engineering	16:20	28.041
Video _070	Mechanical Engineering	09:40	22.407
Video _071	Artificial Intelligence	37:20	233.625
Video _072	Artificial Intelligence	41:16	138.855
Video _073	Artificial Intelligence	01:14:25	328.862
Video _074	Artificial Intelligence	27:12	47.841
Video _075	Artificial Intelligence	50:00	340.788
Video _076	Artificial Intelligence	50:41	342.631
Video _077	Artificial Intelligence	01:16:52	383.113

Table 1 Continued

Video ID	Type of Video	Video length (mm:ss)	Video size (KB)
Video _078	Artificial Intelligence	09:50	39.436
Video _079	Artificial Intelligence	01:06:05	112.198
Video _080	Artificial Intelligence	32:39	53.107
Video _081	Database	01:07:28	149.700
Video _082	Database	01:20:54	176.777
Video _083	Database	01:13:06	161.055
Video _084	Database	01:10:32	219.854
Video _085	Database	01:21:11	232.809
Video _086	Database	01:13:38	226.143
Video _087	Database	01:17:16	238.939
Video _088	Database	01:17:22	182.382
Video _089	Database	01:08:01	182.484
Video _090	Database	52:13	154.771
Video _091	Algorithms and Data Structure	01:11:04	307.258
Video _092	Algorithms and Data Structure	01:11:54	341.530
Video _093	Algorithms and Data Structure	01:14:58	324.127
Video _094	Algorithms and Data Structure	01:07:21	293.809
Video _095	Algorithms and Data Structure	01:06:17	257.611
Video _096	Algorithms and Data Structure	01:07:54	309.053
Video _097	Algorithms and Data Structure	07:22	12.408
Video _098	Algorithms and Data Structure	09:58	16.762
Video _099	Algorithms and Data Structure	13:17	18.532

Table 1 Continued

Video ID	Type of Video	Video length (mm:ss)	Video size (KB)
Video _100	Algorithms and Data Structure	06:14	10.789
Video _101	Networking	09:09	18.252
Video _102	Networking	10:48	65.470
Video _103	Networking	13:14	24.898
Video _104	Networking	07:14	23.727
Video _105	Networking	07:05	19.993
Video _106	Networking	56:47	129.332
Video _107	Networking	32:50	41.539
Video _108	Networking	06:08	10.484
Video _109	Networking	22:30	41.773
Video _110	Networking	10:19	25.218

Today, all lecture videos are shot in different environmental conditions. For this reason, all videos in the dataset were selected from different environmental conditions. With this method, it has been tried to obtain results close to real life comparisons.

The levels of the videos are displayed in the preceding table. As depicted in the diagram, the dataset is evaluated at three distinct levels during the research.

For the level 1 research, the dataset consists of education, medicine, social and engineering modules, and the total of videos is 110. There are 10 videos are in education, 10 videos in medicine, 10 videos in social, and 80 videos in engineering modules. In this level of videos consist of lecture videos.

For the level 2 research, the dataset consists of electronic engineering, mechanical engineering and computer engineering modules, and the total of videos is 80. There are 20 videos are in Electronic Engineering, 20 videos in Mechanical Engineering, and 40 videos in Computer Engineering Modules. In this level of videos consist of Engineering videos.

For the level three research, our dataset consists of Artificial Intelligence, Database, Networking, Algorithms and Data Structure and Engineering modules, and the total of videos is 40. There are 10 videos are in Artificial Intelligence, 10 videos in

Database, 10 videos in Networking, and 10 videos in Algorithms and Data Structure and Engineering Modules. In this level of videos consist of Computer Engineering videos.

Table 2: Distribution of dataset by levels and classes

Level 1	**Level 2**	**Level 3**	**Number of Video**
Education	-	-	10
Medicine	-	-	10
Social	-	-	10
Engineering	Electronic Engineering	-	20
	Mechanical Engineering	-	20
	Computer Engineering	Artificial Intelligence	10
		Database	10
		Algorithms and Data Structure	10
		Networking	10

As depicted in the first diagram, the length and size of the videos are different. The primary aim of building a dataset consisting of distinct classes at different levels is to determine whether algorithms differ in different semantic contexts and which algorithms produce superior outcomes in which semantic contexts. Our dataset consists of different types of videos in order to make a better analysis. For example, I would like to point out that there is no speech in the 105th video. For this reason, we don't expect 100 percent accuracy for the ASR module. Detailed information about the dataset can be found in Table 2.

CHAPTER IV
VIDEO INDEXING SYSTEMS

Video indexing is the technique of collecting data from videos to aid users in locating and gaining access to videos. It is derived from the phrase indexing, which generally refers to making information more accessible and presentable.

Two technology of video indexing systems ASR and OCR are taken into account.

4.1 AUTOMATIC SPEECH RECOGNITION (ASR)

ASR is the technique of obtaining the transcription of speech, often known as a word sequence, by analyzing the shape of the speech wave [17]. Natural Language Processing (NLP) is the foundation of the most advanced form of ASR technologies now under development. This variant of ASR comes the closest to enabling human-machine conversation, and while it still has a long way to go before reaching its full potential, we're already seeing remarkable results in the form of intelligent smartphone interfaces such as the Siri app on the iPhone and other systems used in business and advanced technology contexts [18]. The architecture of ASR is given in Figure 1.

In order to stream Speech to text in the analysis part, the REV application programming interface (API) is used https://docs.rev.ai/api/streaming/api.

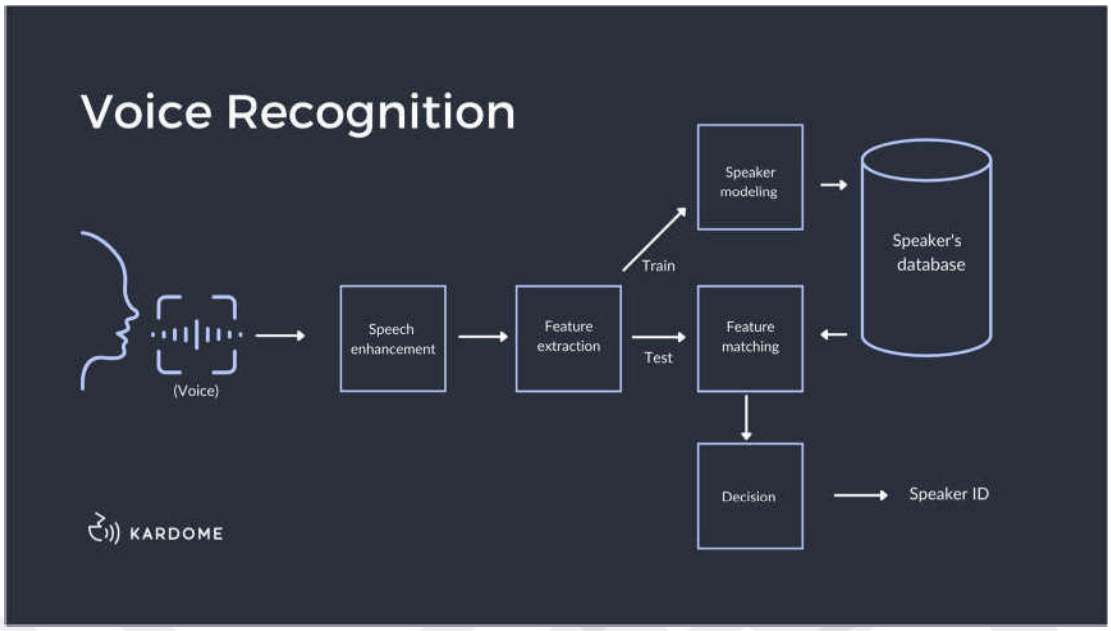

Figure 1: ASR architecture [19]

4.2 OPTICAL CHARACTER RECOGNITION (OCR)

OCR is the abbreviation for optical character recognition. This technology enables automatic character recognition via an optical system. Regarding the anatomy of the human body, the eyes are optical mechanisms. The image viewed by the eyes serves as input for the brain. The ability to comprehend these inputs differs among individuals based on a variety of criteria. OCR is a technology that mimics human reading ability [20]. The architecture of OCR is given in Figure 2.

OCR is a technology that enables the conversion of various document types, such as scanned paper documents, PDF files, and digital camera photos, into editable and searchable data. The photos captured by a digital camera differ from the scanned image or document. They frequently contain flaws such as edge distortion and weak lighting, making it challenging for the majority of OCR programs to effectively recognize text. Tesseract can be chosen due to its general acceptance, extensibility and flexibility, active development community, and "out of the box" functionality [20].

For textual data extraction [21], tesseract OCR software is employed. Moreover, Google is now responsible for the development of the tesseract OCR software. It is among the most precise open-source tools for OCR. Every 30 frames, or once per second, OCR outputs are captured from all videos. Extraction of textual data is performed every second. Transitions between slides or text that last less than a second are disregarded. Since this topic had already been explored in Ağzıyağlı's [5]

thesis, only the OCR results from his thesis are used when OCR and ASR approaches are combined.

Figure 2: OCR architecture [22]

CHAPTER V
CLASSIFICATION METHODS

5.1. NAÏVE BAYES

Thomas Bayes invented the conditional probability calculation formula in 1812 [23]. This theorem illustrates the relationship between conditional probabilities and marginal probabilities inside a random variable's probability distribution.

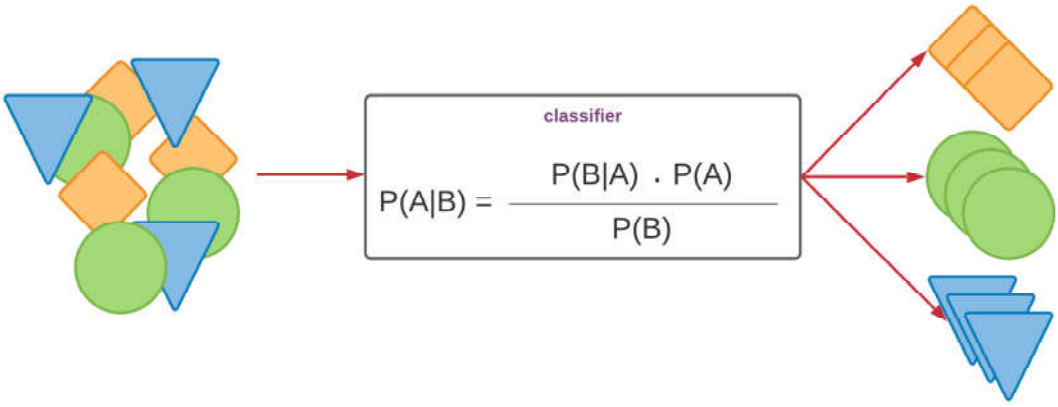

Figure 3: Naïve bayes classifier formula [24]

P (A|B) = B probability of event A occurring

P (A) = The probability of event A occurring

P (B| A) = The probability of event B occurring when event A occurs

P (B) = The probability of event B occurring [25]

The Naïve Bayes classifier utilizes Bayes' theorem. It is a "lazy" learning method capable of operating on unstable datasets. The method calculates the likelihood of each state for each element and classifies them according to the state with the highest probability. With minimal training data, it is able to produce highly successful results. The formula of the Bayes theorem is given in Figure 3.

Advantages

- Training requires minimal computational time. [26].
- It has excellent performance [26].

- It improves the performance of categorization by deleting irrelevant features [26].

Disadvantages

- The Naïve Bayes classifier requires a massive number of records to produce accurate results [26].
- Less accurate than other classifiers on some datasets [26].

5.2 SUPPORT VECTOR MACHINE

SVM was originally mentioned in 1992 when Boser, Guyon, and Vapnik has described it in COLT-92. SVMs are a group of related methods for classification and regression based on supervised learning. They are related to generalized linear classifiers. SVM is an alternative method for a classification and regression prediction tool that utilizes machine learning theory to enhance predicted accuracy while systematically omitting overfitting the data. SVMs use the hypothesis space of linear functions in a high-dimensional feature space and are taught with an optimization-based learning algorithm that carries out a statistical learning bias. It is appropriate for complex, small to medium scale datasets [27]. The formula of the SVM classifier is given in Figure 4.

As can be seen in the formula;

$\hat{y} = 0$ if $w^T * x + b < 0$

$\hat{y} = 1$ if $w^T * x + b >= 0$

w; weight vector,

x; input vector,

b; bias ($\theta 0$)

5.3 RANDOM FOREST

As a general-purpose classification and regression technique, the Random Forest algorithm, developed Breiman in 2001, has proven to be a highly effective technique. The strategy, which mixes many randomized decision trees and averages their predictions, has demonstrated exceptional performance in circumstances with a large number of variables and observations. In addition, it is flexible enough to be applied to large-scale problems, easily adaptable to a variety of ad-hoc learning

activities and returns metrics of varying relevance [7]. The technique of the Random Forest is given in Figure 5.

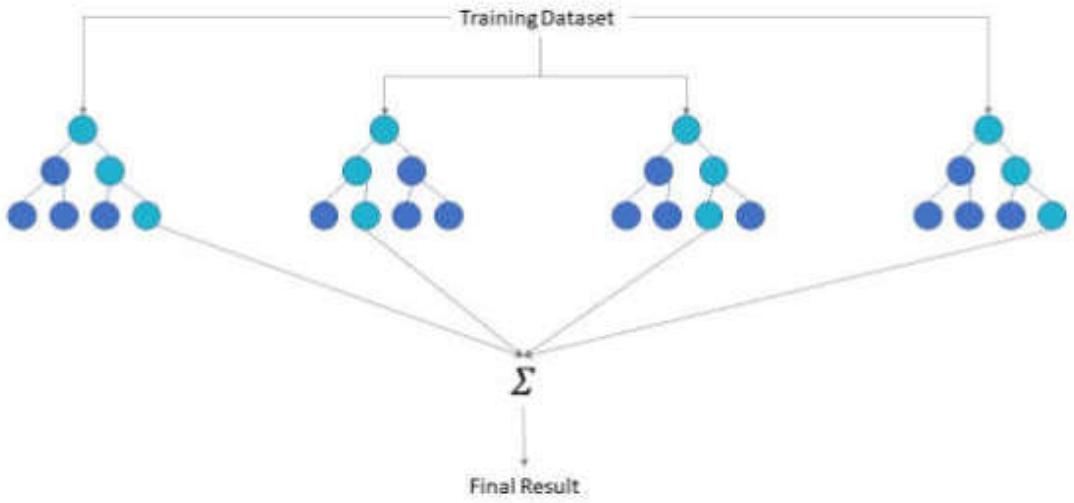

Figure 4: Random Forest Technique [28]

Before training the dataset, three primary hyperparameters must be established for Random Forest algorithms. Included are node size, the number of trees, and the number of features sampled. The Random Forest classifier can then be applied to regression or classification issues [29].

The Random Forest technique contains a series of decision trees, and each tree in the ensemble is comprised of a bootstrap sample. The bootstrap sample is a data sample retrieved from a training set with substitute. One-third of the training sample is set aside as test data, also known as the out-of-bag sample, which is mentioned again throughout the thesis. Then, feature bagging introduces another instance of randomness, increasing the diversity of the dataset and decreasing the correlation between decision trees. The determination of the prediction varies based on the nature of the situation. For a regression job, the individual decision trees are averaged, though, for a classification task, the predicted class is established by the majority vote of the most common categorical variable. Finally, the out-of-bag sample is utilized for cross-validation to conclude the prediction [2].

CHAPTER VI
METHODOLOGY

6.1 ARCHITECTURE DESIGN

Using a dataset consisting of 110 lecture videos, this thesis classifies videos via classical classification techniques. OCR and ASR video indexing systems are utilized for this purpose. After receiving the textual data, three distinct classification techniques are used to categorize it.

Textual information is extracted from the images using OCR and ASR outputs from the videos. The sections "6.3 Extraction of textual data from video with Automatic Speech Recognition", "6.4 Extraction of textual data from video with Optical Character Recognition", and "6.5 Extraction of textual data from video with Automatic Speech Recognition and Optical Character Recognition" contain additional information on textual data extraction.

Depending on the classification methods, sequential preprocesses are applied to the textual data extracted following the OCR and ASR outputs in order to make them appropriate for classification. All textual data are analyzed using three distinct classification algorithms based on classic machine learning techniques. This thesis employs the traditional machine learning methods of Naïve Bayes, SVM, and Random Forest.

For conventional machine learning techniques, preprocesses have included converting the data to lowercase letters and removing extraneous words, checking the words against the English lexicon, determining the word frequency, normalizing the word frequency, and converting the data to the ‘. arff' format. Similarly, the study is conducted without normalizing the word frequency, and comparisons are done with the normalized results. Additional information regarding data preprocessing for traditional machine algorithms "6.6 Data preprocessing for Naïve Bayes, Support Vector Machine and Random Forest." Traditional machine learning algorithms generate frequency vectors through preprocessing. The classification which has been made with traditional machine learning methods is carried out in the Weka program.

This overall architecture is replicated for three distinct video indexing systems. In the initial phase, text files comprising OCR following the ASR approach and the result of the final two are analyzed. In the results section, these three distinct strategies are contrasted and analyzed using a number of different metrics. The general architecture of the system and detailed information can be found in Figure 6.

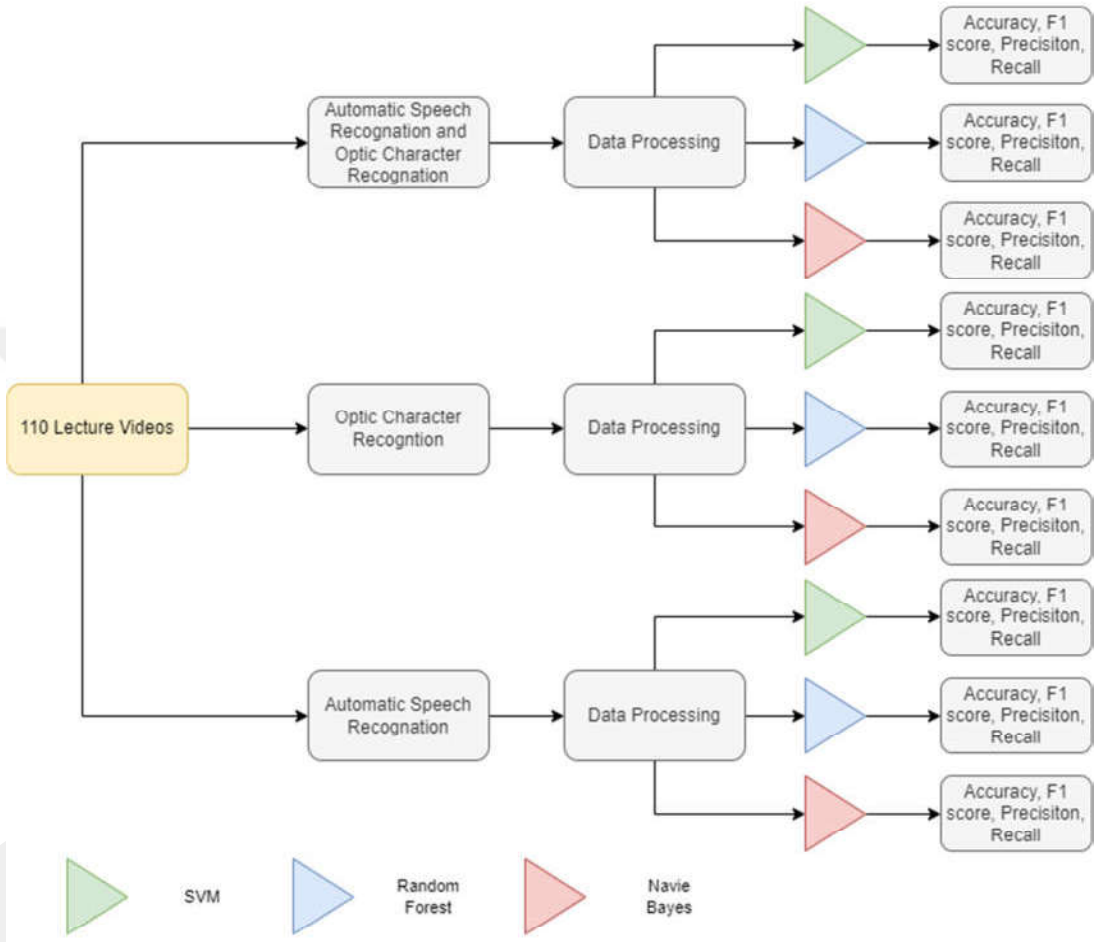

Figure 5: General Architectural Flow Chart

6.2 EXTRACTION OF TEXTUAL DATA FROM VIDEO WITH AUTOMATIC SPEECH RECOGNITION

In this section, ASR is used to convert 110 lecture videos into audio files. This is accomplished by minimizing the size of the videos prior to their conversion to text. Thus, it is them into text videos with more speed and efficiency. After then, the speech-to-text API available at "https://www.rev.ai/jobs/speech-to-text" is implemented. Thus, 110 lecture videos are transformed into text videos.

In this thesis, the main language of all training videos is English. But Rev Api also supports Arabic, Danish, Dutch, English, Persian, French, German, Hebrew,

Hindi, Indonesian, Italian, Japanese, Korean, Malay, Mandarin, Russian, Spanish, Tamil, Telugu, Turkish. Moreover, Rev Api provides noise suppression for all of the videos one by one.

6.3 EXTRACTION OF TEXTUAL DATA FROM VIDEO WITH OPTICAL CHARACTER RECOGNITION

The software for OCR (Tesseract) is used to extract textual data. According to the Dataset section, the frame rate of the used videos is 30 per second. All Tesseract outputs are captured once in 30 frames for this thesis. Once per second, textual data extraction is performed. The absence of sub-second slides or text transitions. In this section, since the same dataset is used for analysis, the results from Ağzıyağlı's thesis [5] are utilized directly. An examination of Ağzıyağlı's thesis [5] is suggested for more information about OCR and what has been accomplished in this section.

6.4 EXTRACTION OF TEXTUAL DATA FROM VIDEO WITH AUTOMATIC SPEECH RECOGNITION AND OPTICAL CHARACTER RECOGNITION

Using ASR and OCR, it is combined the text files extracted from the videos in this section. In future analyses, it can be utilized a combination of ASR and OCR text datasets. The extraction of the textual data flow chart can be found in Figure 7.

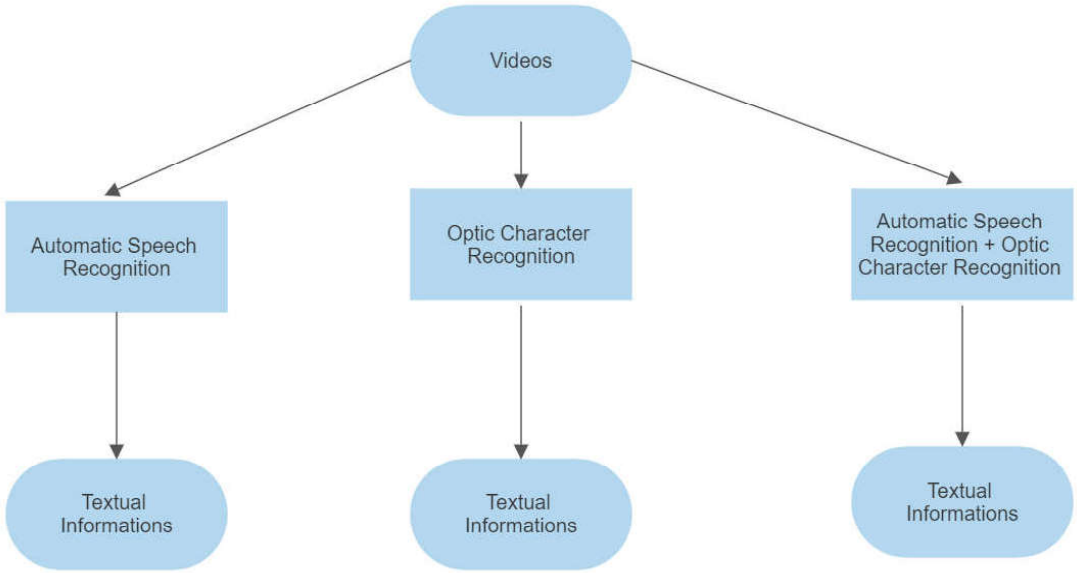

Figure 6: Extraction of textual data flow chart

6.5 DATA PREPROCESSING FOR NAÏVE BAYES, SUPPORT VECTOR MACHINES, AND RANDOM FOREST

There are 3 different textual information obtained from 3 different methods as ASR, OCR, and ASR+OCR. During the data preprocessing phase, various preprocessing steps are performed on the obtained text files in order to apply the Naïve Bayes, SVM, and Random Forest algorithms. In its broadest sense, the process consists of converting the textual data from the lecture videos into frequency vectors for the application of the Naïve Bayes, SVM, and Random Forest algorithms.

The textual data obtained from these three methods are initially converted to lowercase letters. Although the case of a word has little significance in terms of semantics, they have distinct values in computer architecture. It depends on various ASCII codes, and this variation may manifest as distinct characteristics in the classification section.

Second, unnecessary words are eliminated. Internet-accessible, open-source "Stop Word" lists are utilized here. In this context, "unnecessary words" refers to those that lack significant meaning in an English sentence. These unnecessary words can be eliminated from sentences without sacrificing any meaning. These unnecessary words have been eliminated.

In the third step, it is determined if the remaining words are included in the English dictionary. Words that are not found in the English dictionary, are eliminated. Some of Tesseract's words are meaningless, including non-English words. At this point, such a check is performed, and meaningless words are eliminated.

After these steps, the remaining words are read to create a word pool. The similarity ratio is the primary factor that determines the number of elements in the word pool. After applying the entire flow, it is determined which words appeared in which videos and classifications are made based on this information. For each piece of text data, processes such as removing unnecessary words and omitting non-English, meaningless words are applied to reduce the word pool. After creating the word pool, the frequency of each word in the video is determined by counting the number of times it appeared in the video. The frequency is then normalized based on the total number of words, and the analysis is conducted without normalization. The outcomes are evaluated in these two ways. The data processing for traditional algorithms and detailed information can be found in Figure 8.

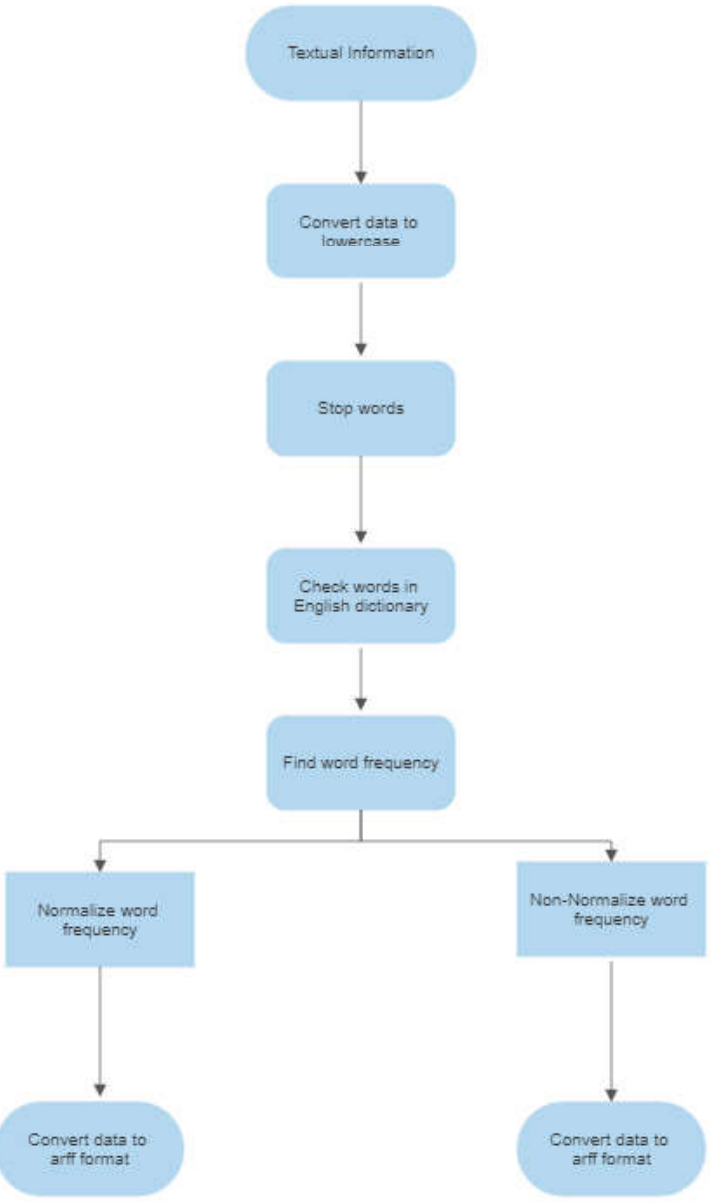

Figure 7: Data processing for Traditional Algorithms

6.6 CLASSIFICATION BY MEANS OF NAÏVE BAYES, SUPPORT VECTOR MACHINES, AND RANDOM FOREST

Utilized data at this point are output files containing the frequency of each word in the video. It is used the Weka application to classify these output files. Consequently, the datasets are prepared based on the type of the "Weka Dataset".

Weka supports ".arff" files extracted during the data preprocessing phase for Naïve Bayes, SVM, and Random Forest. Not only as a format but also as a sequence of features, it is structured under the classification in the Weka machine learning program.

The Weka workbench is comprised of graphical user interfaces and a series of visualization tools and algorithms for data analysis and predictive modeling. The Weka software is an open-source tool. It is portable and platform-independent due to its complete implementation in the Java programming language, which enables it to run on virtually every modern computing platform [30].

In addition to that, classification with Weka is carried out on a total of six different datasets. The outputs of ASR and ASR+OCR that have been normalized, as well as the outputs that have not been normalized, are presented here.

CHAPTER VII
EXPERIMENT AND RESULTS

As indicated in the dataset section, four video lectures are available for Level 1: education, medical, social, and engineering. There are 110 total videos at this level. Level 2 video courses include: electronics and electronics engineering, mechanical engineering, and computer engineering. There are 80 total videos at this level. Level 3 video courses include: artificial intelligence, database, algorithms and data structure and networking. There are a total of 40 videos at this level.

During the experiments, accuracy, F1 score, precision and recall values are calculated separately for these 3 levels. These 4 different metrics are analyzed using traditional machine learning algorithms, Naïve Bayes, Random Forest and SVM.

7.1 AUTOMATIC SPEECH RECOGNITION RESULTS

As a result of the ASR, a total of 14512 unique words are recovered from the first-level classification dataset, which consisted of 110 lecture videos. A total of 11153 unique words are recovered from the second-level classification dataset, which consisted of 80 lecture videos. A total of 8625 unique words are recovered from the first level classification dataset, which consisted of 40 lecture videos. In Figure 9, the number of unique words in the analysis of ASR is given for each level.

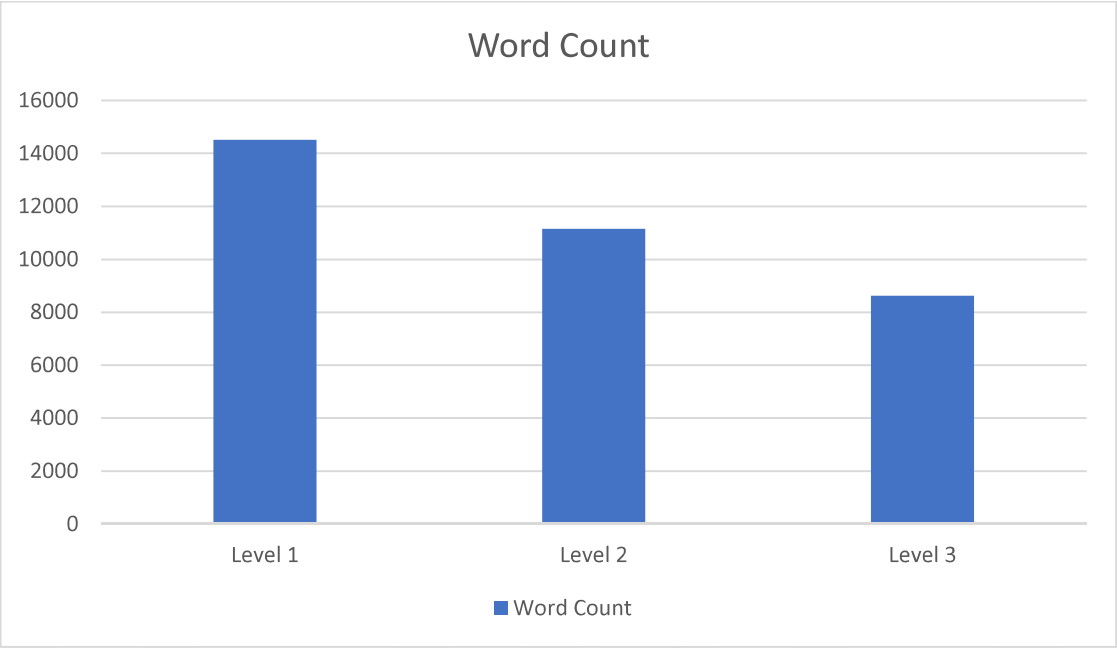

Figure 8: Unique word counts for ASR

7.1.1 Accuracy Values

Naïve Bayes has provided an accuracy value of 32.7273%, Random Forest has provided a value of 72.7273%, and SVM has provided a value of 91.8182% in the normalized analysis in level 1. In the unnormalized analysis at level 1, Naïve Bayes has given an accuracy value of 91.8182%, Random Forest has given a value of 72.7273% and SVM has given a value of 91.8182%.

Naïve Bayes has provided an accuracy value of 41.25%, Random Forest has provided a value of 86,25%, and SVM has provided a value of 91.25% in the normalized analysis in level 2. In the unnormalized analysis at level 2, Naïve Bayes has given an accuracy value of 92.50%, Random Forest has given a value of 82.50% and SVM has given a value of 75%.

Naïve Bayes has provided an accuracy value of 60%, Random Forest has provided a value of 87.5%, and SVM has provided a value of 97.50% in the normalized analysis in level 3. In the unnormalized analysis at level 3, Naïve Bayes has given an accuracy value of 92.5%, Random Forest has given a value of 82.50% and SVM has given a value of 85%.

Normalized accuracy values obtained by traditional machine learning methods for ASR can be found in Figure 10.

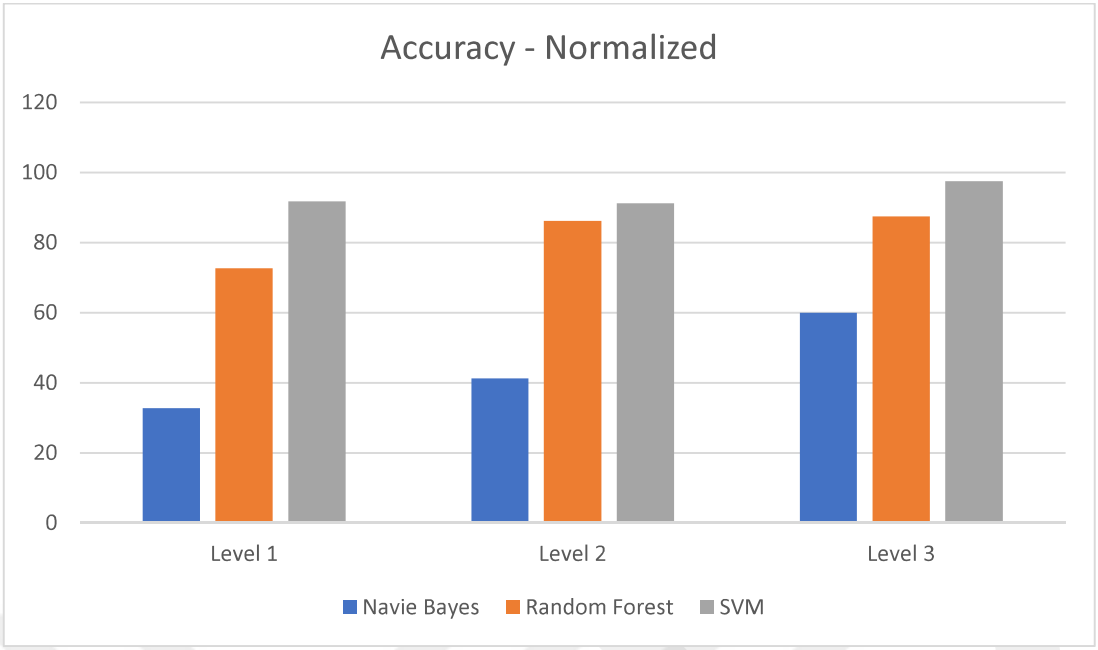

Figure 9: Normalized accuracy values obtained by traditional machine learning methods for ASR

Unnormalized accuracy values obtained by traditional machine learning methods for ASR can be found in Figure 11.

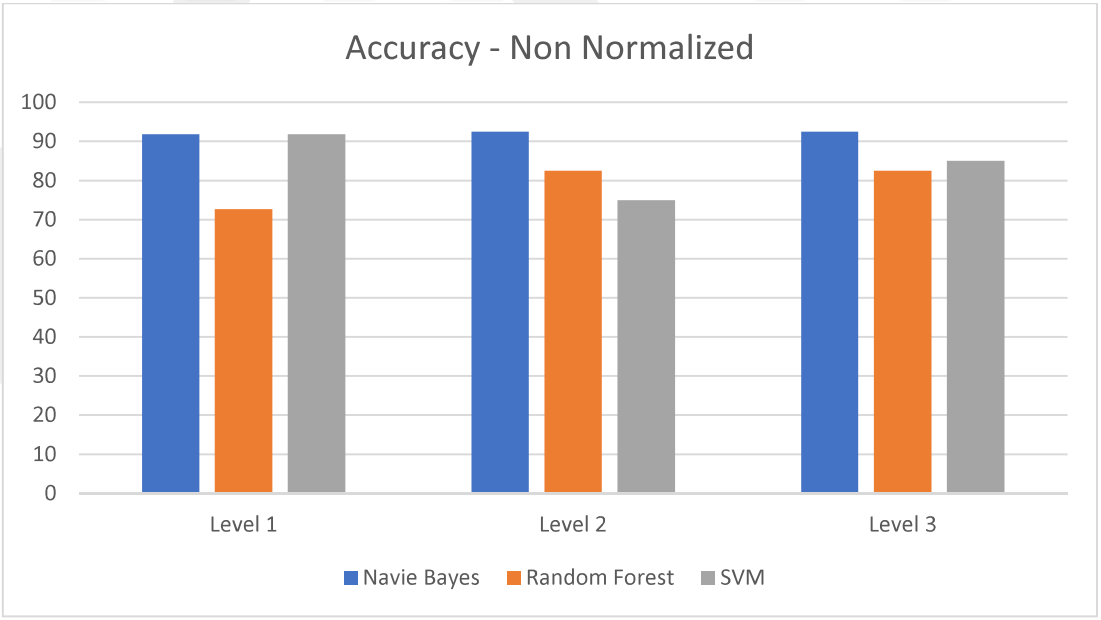

Figure 10: Unnormalized accuracy values obtained by traditional machine learning methods for ASR

7.1.2 F1 score Values

Naïve Bayes and Random Forest have yielded no results in level 1 normalized analysis. The SVM has given a score of 0.912 as an F1 score. In the unnormalized analysis at level 1, values are not obtained for Random Forest, while values of 0.919 and 0.912 are obtained for Naïve Bayes.

Naïve Bayes has yielded no results in level 2 normalized analysis. The SVM has given a score of 0.912. Random Forest has given a score of 0.861. In the unnormalized analysis at level 2, Naïve Bayes has given an F1 score value of 0.925, Random Forest has given a value of 0.82 and SVM has given a value of 0.756.

Random Forest has provided a value of 0.876, and SVM has provided a value of 0,975 in the normalized analysis in level 3. In the unnormalized analysis at level 3, Naïve Bayes has given an F1 score value of 0.924, Random Forest has given a value of 0,829 and SVM has given a value of 0,854.

Normalized F1 score values obtained by traditional machine learning methods for ASR can be found in Figure 12.

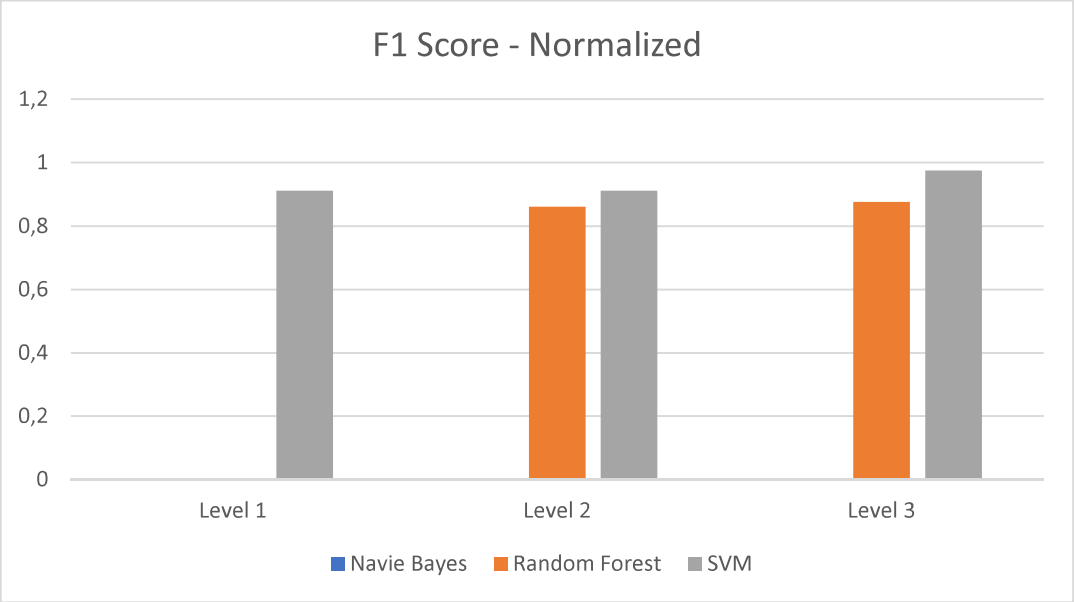

Figure 11: Normalized F1 score values obtained by traditional machine learning methods for ASR

Unnormalized F1 score values obtained by traditional machine learning methods for ASR can be found in Figure 13.

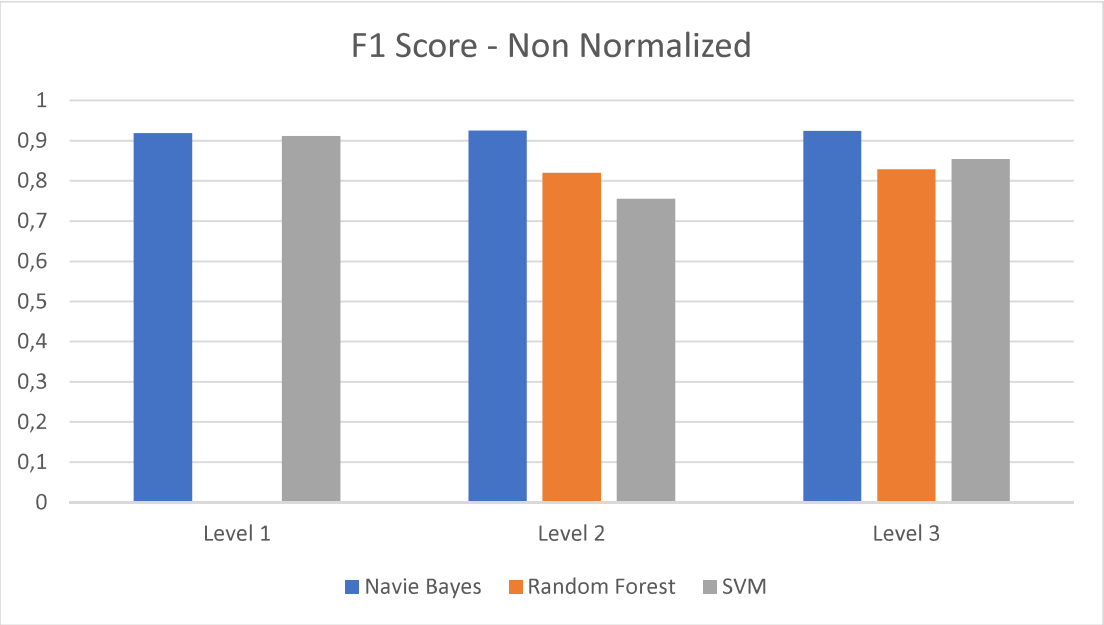

Figure 12: Unnormalized F1 score values obtained by traditional machine learning methods for ASR

7.1.3 Precision Values

Naïve Bayes and Random Forest have yielded no results in level 1 normalized analysis. The SVM has given a score of 0.921 as a precision value. In the unnormalized analysis at level 1, values are not obtained for SVM, while values of 0.921 and 0.919 are obtained for Naïve Bayes.

Random Forest has provided a value of 0.874 and SVM has provided a value of 0.919 in the normalized analysis in level 2. In the unnormalized analysis at level 2, Naïve Bayes has given a precision value of 0.925, Random Forest has given a value of 0.843 and SVM has given a value of 0.772.

Random Forest has provided a value of 0.904, and SVM has provided a value of 0,977 in the normalized analysis in level 3. In the unnormalized analysis at level 3, Naïve Bayes has given a precision value of 0.936, Random Forest has given a value of 0,852 and SVM has given a value of 0,885.

Normalized precision values obtained by traditional machine learning methods for ASR can be found in Figure 14.

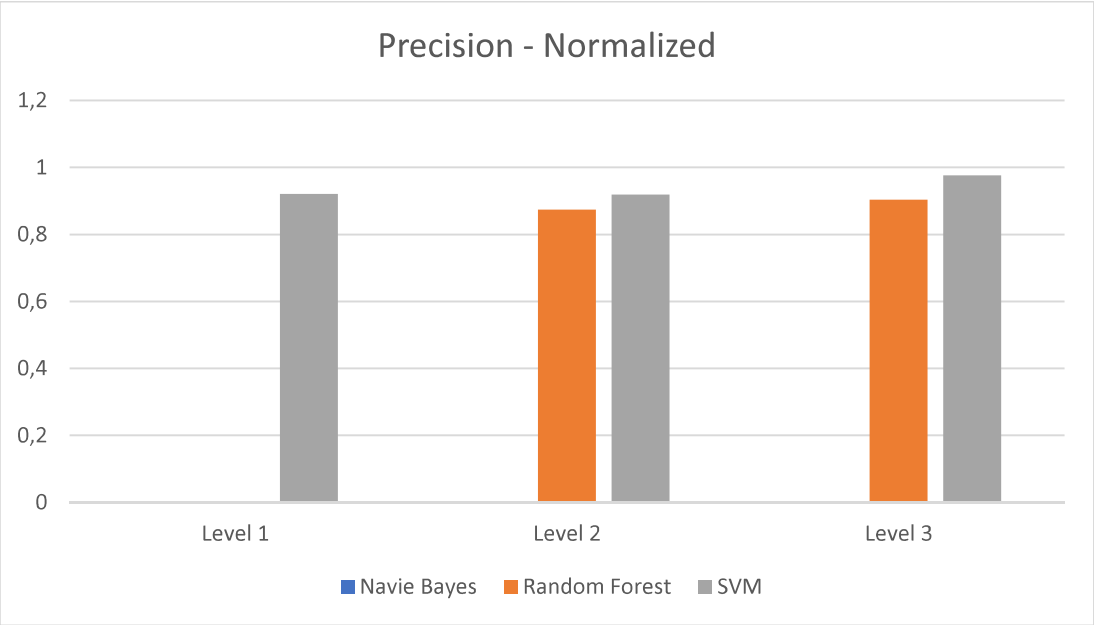

Figure 13: Normalized precision values obtained by traditional machine learning methods for ASR

Unnormalized precision values obtained by traditional machine learning methods for ASR can be found in Figure 15.

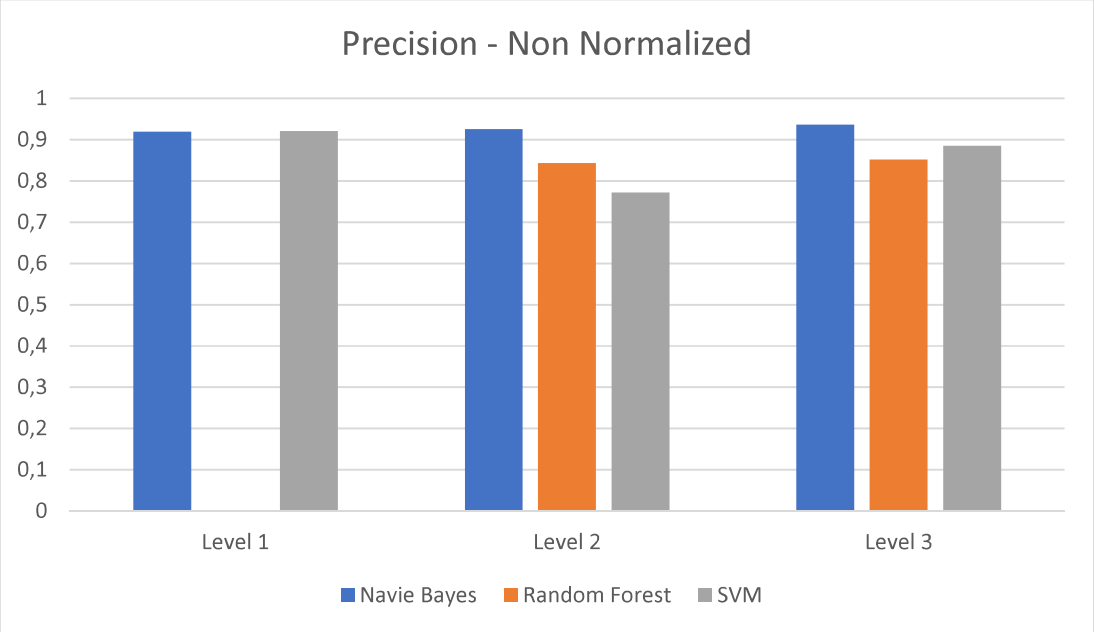

Figure 14: Unnormalized precision values obtained by traditional machine learning methods for ASR

7.1.4 Recall Values

Naïve Bayes has provided a recall value of 0.973, Random Forest has provided a value of 0.727, and SVM has provided a value of 0.918 in the normalized analysis in level 1. In the unnormalized analysis at level 1, Naïve Bayes has given a recall value of 0.918, Random Forest has given a value of 0.7227 and SVM has given a value of 0.918.

Naïve Bayes has provided a recall value of 0.943, Random Forest has provided a value of 0.863, and SVM has provided a value of 0.913 in the normalized analysis in level 2. In the unnormalized analysis at level 2, Naïve Bayes has given a recall value of 0.9250, Random Forest has given a value of 0.825 and SVM gave a value of 0.75.

Naïve Bayes has provided a recall value of 0.889, Random Forest has provided a value of 0.875, and SVM has provided a value of 0.975 in the normalized analysis in level 3. In the unnormalized analysis at level 3, Naïve Bayes has given a recall value of 0.925, Random Forest has given a value of 0.825 and SVM has given a value of 0.85.

Normalized recall values obtained by traditional machine learning methods for ASR can be found in Figure 16.

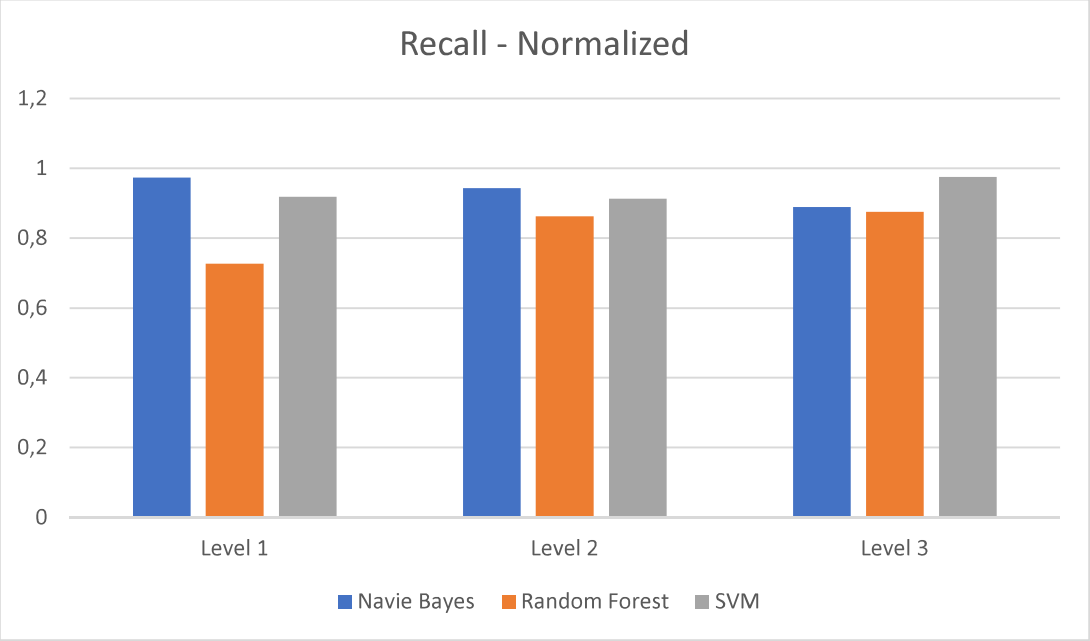

Figure 15: Normalized recall values obtained by traditional machine learning methods for ASR

Unnormalized recall values obtained by traditional machine learning methods for ASR can be found in Figure 17.

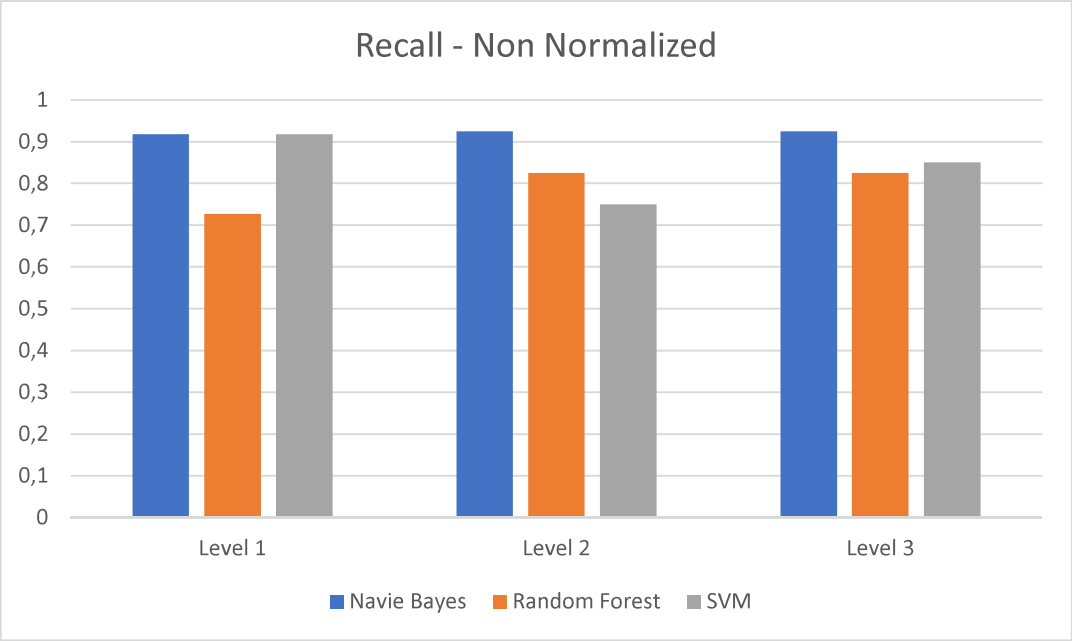

Figure 16: Unnormalized recall values obtained by traditional machine learning methods for ASR

7.2 AUTOMATIC SPEECH RECOGNITION AND OPTICAL CHARACTER RECOGNITION RESULTS

As a result of the ASR and OCR, A total of 91978 unique words are recovered from the first level classification dataset, which consisted of 110 lecture videos, and a total of 73912 unique words are recovered from the second level classification dataset, which consisted of 80 lecture videos and a total of 62239 unique words are recovered from the first level classification dataset, which consisted of 40 lecture videos. In Figure 18, the number of unique words in the analysis of ASR and OCR is given for each level.

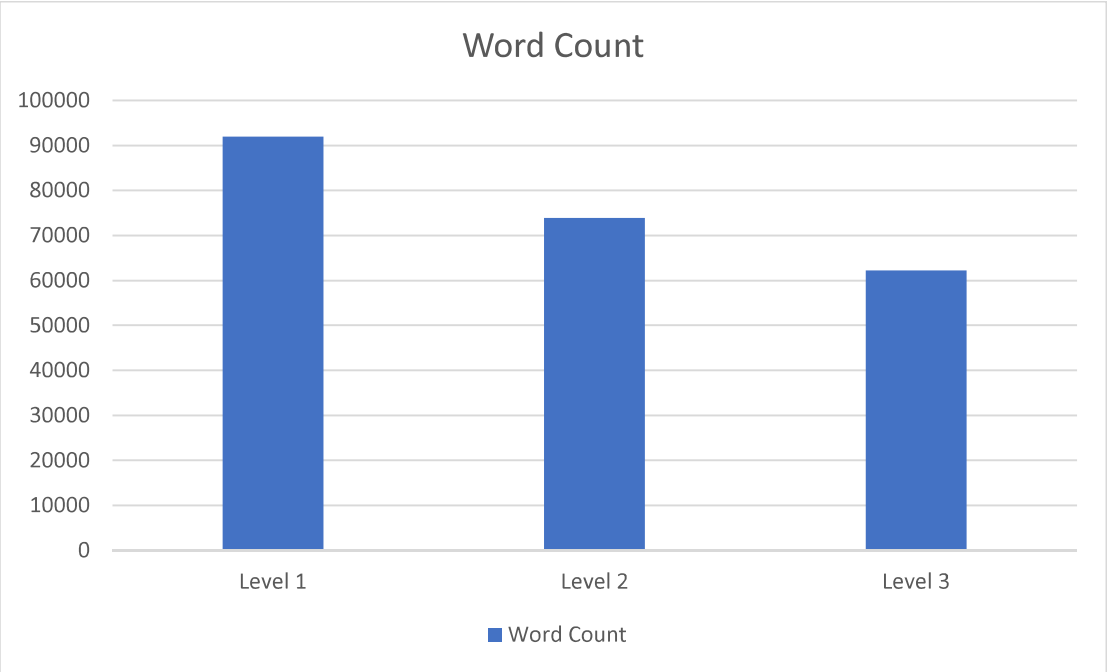

Figure 17: Unique word counts for ASR and OCR

7.2.1 Accuracy Values

Naïve Bayes has provided an accuracy value of 77.2727%, Random Forest has provided a value of 72.7273%, and SVM has provided a value of 85.4545% in the normalized analysis in level 1.

Naïve Bayes has provided an accuracy value of 75%, Random Forest has provided a value of 85%, and SVM has provided a value of 85% in the normalized analysis in level 2.

Naïve Bayes has provided an accuracy value of 72.5%, Random Forest has provided a value of 80%, and SVM has provided a value of 62.50% in the normalized analysis in level 3.

For all 3 levels, the unnormalized case has given almost the same results as the normalized case.

Normalized accuracy values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 19.

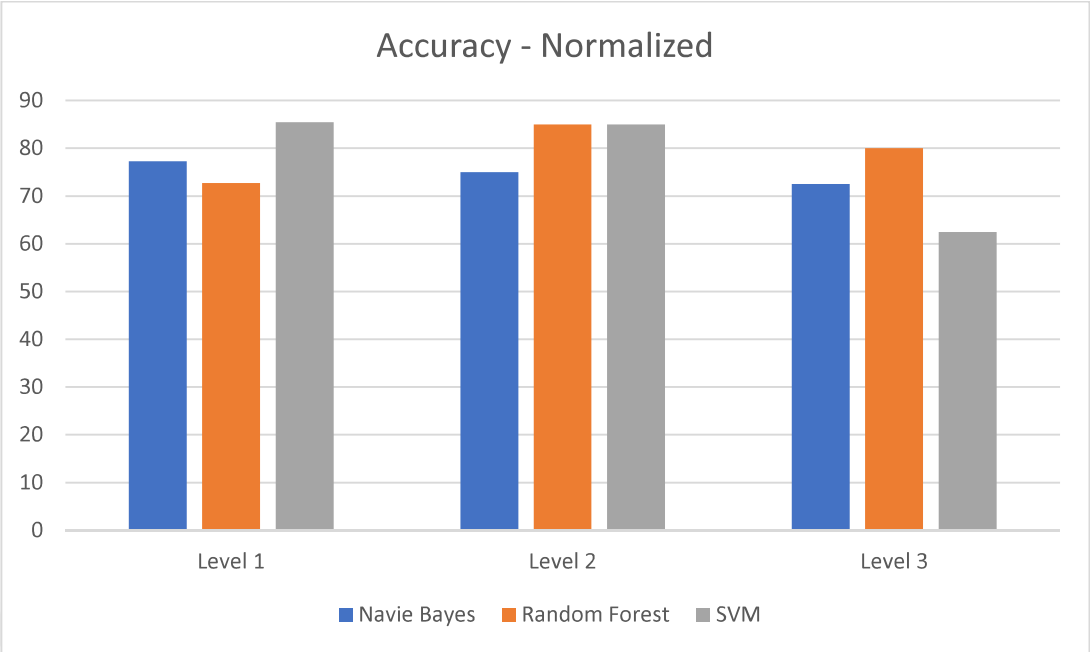

Figure 18: Normalized accuracy values obtained by traditional machine learning methods for ASR and OCR

Unnormalized accuracy values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 20.

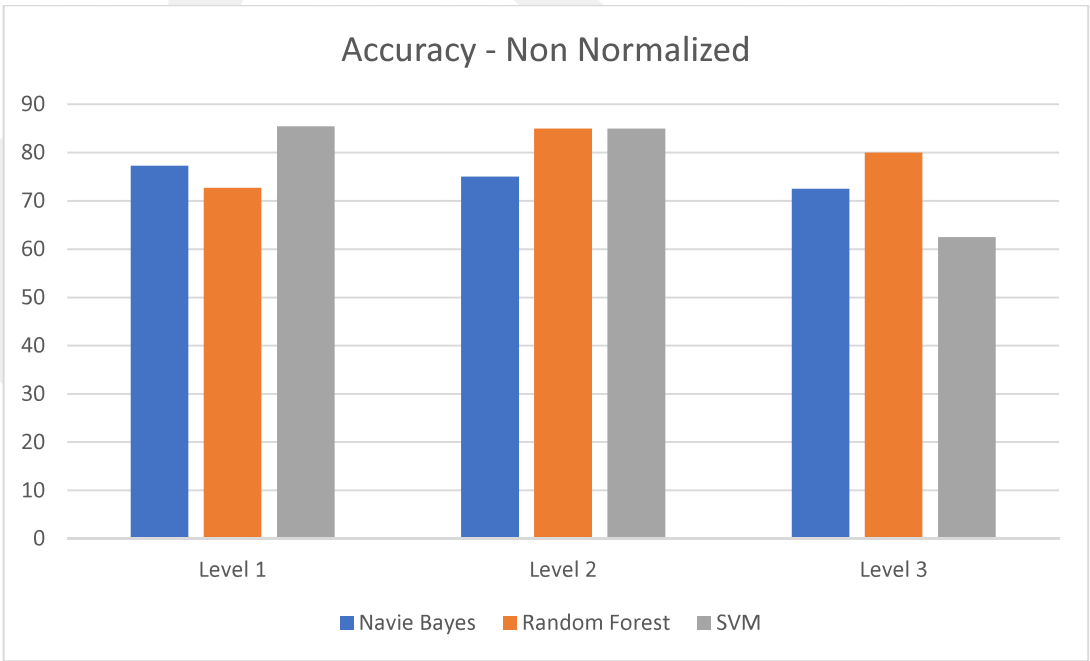

Figure 19: Unnormalized accuracy values obtained by traditional machine learning methods for ASR and OCR

7.2.2 F1 score Values

Random Forest has yielded no results in level 1 normalized analysis. The SVM has given a score of 0.912 as the F1 score. Naïve Bayes has given a score of 0.786 as an F1 score.

Naïve Bayes has provided an F1 score value of 0.739, Random Forest has provided a value of 0.852, and SVM has provided a value of 0.85 in the normalized analysis in level 2.

Naïve Bayes has provided an F1 score value of 0.726, Random Forest has provided a value of 0.799, and SVM has provided a value of 0.61 in the normalized analysis in level 3.

There is no significant difference between normalized and unnormalized findings.

Normalized F1 score values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 21.

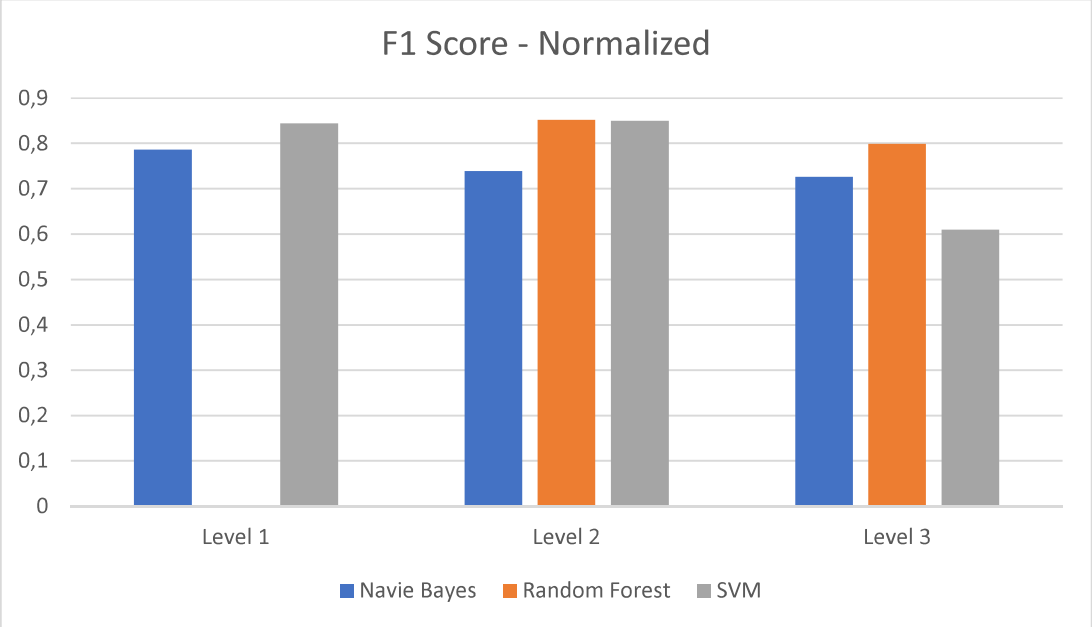

Figure 20: Normalized F1 score values obtained by traditional machine learning methods for ASR and OCR

Unnormalized F1 score values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 22.

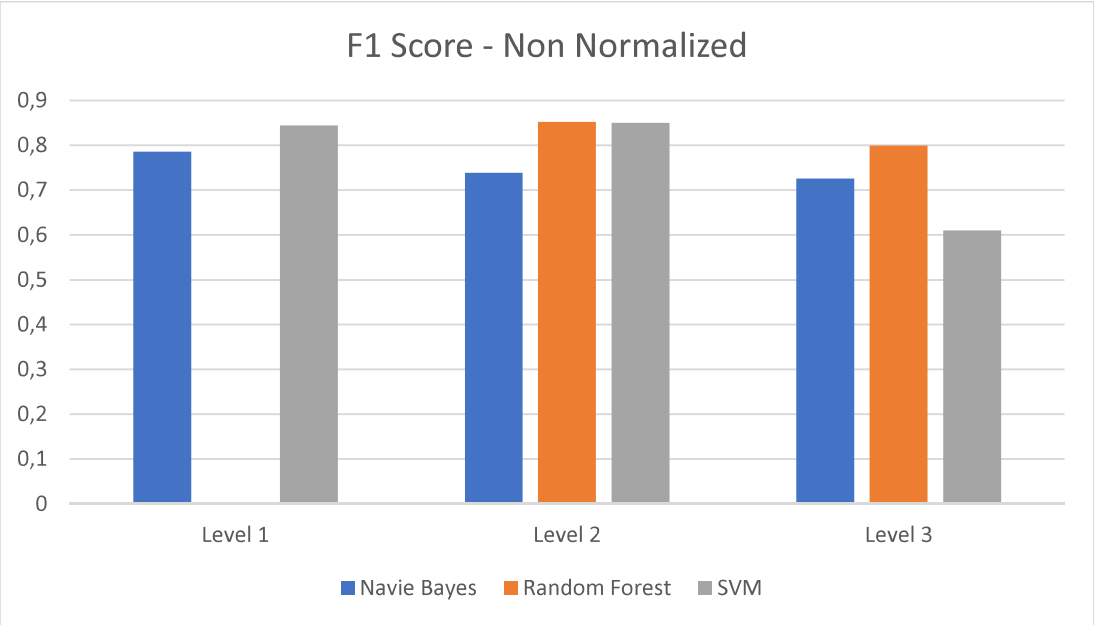

Figure 21: Unnormalized F1 score values obtained by traditional machine learning methods for ASR and OCR

7.2.3 Precision Values

Random Forest has yielded no results in level 1 normalized analysis. Naïve Bayes has given a score of 0.811 as the precision value. The SVM has given a score of 0.874 as a precision value.

Naïve Bayes has provided a precision value of 0.757, Random Forest has provided a value of 0.89 and SVM has provided a value of 0.853 in the normalized analysis in level 2.

Naïve Bayes has provided an F1 score value of 0.742, Random Forest has provided a value of 0.869, and SVM has provided a value of 0,677 in the normalized analysis in level 3.

There is no significant difference between normalized and unnormalized findings.

Normalized precision values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 23.

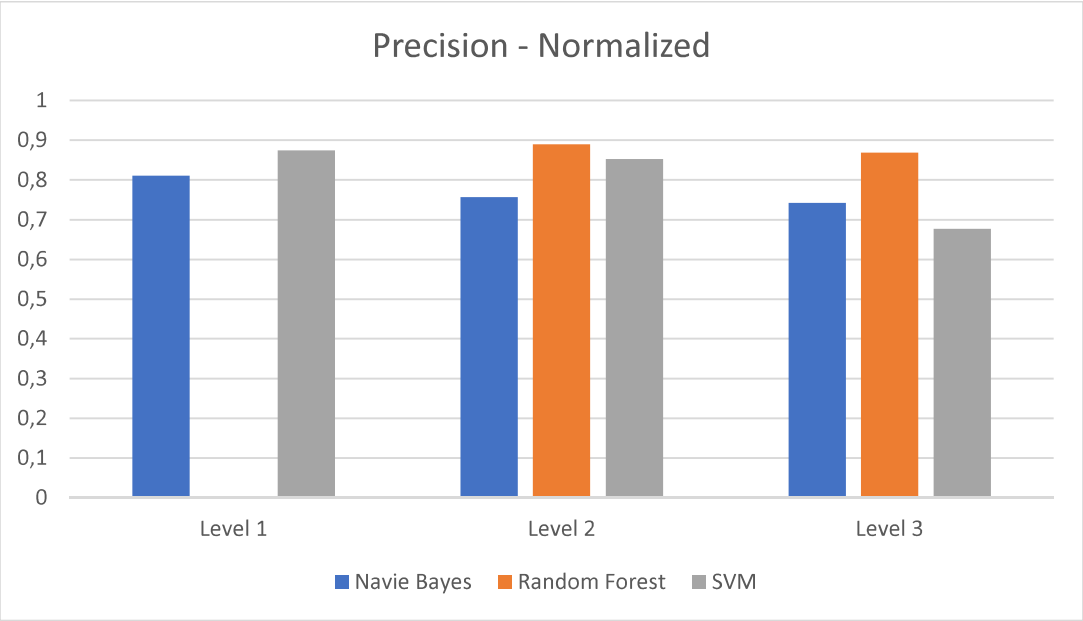

Figure 22: Normalized precision values obtained by traditional machine learning methods for ASR and OCR

Unnormalized precision values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 24.

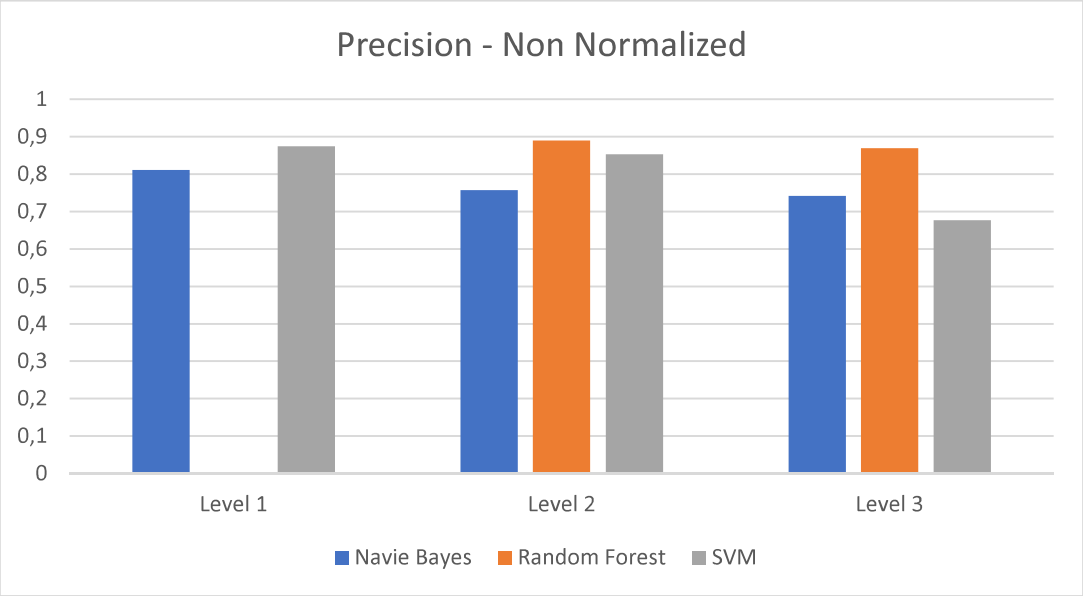

Figure 23: Unnormalized precision values obtained by traditional machine learning methods for ASR and OCR

7.2.4 Recall Values

Naïve Bayes has provided a recall value of 0.773, Random Forest has provided a value of 0.727, and SVM has provided a value of 0.855 in the normalized analysis in level 1.

Naïve Bayes has provided a recall value of 0.75, Random Forest has provided a value of 0.85, and SVM has provided a value of 0.85 in the normalized analysis in level 2.

Naïve Bayes has provided a recall value of 0.725, Random Forest has provided a value of 0.8, and SVM has provided a value of 0.625 in the normalized analysis in level 3.

Normalized recall values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 25.

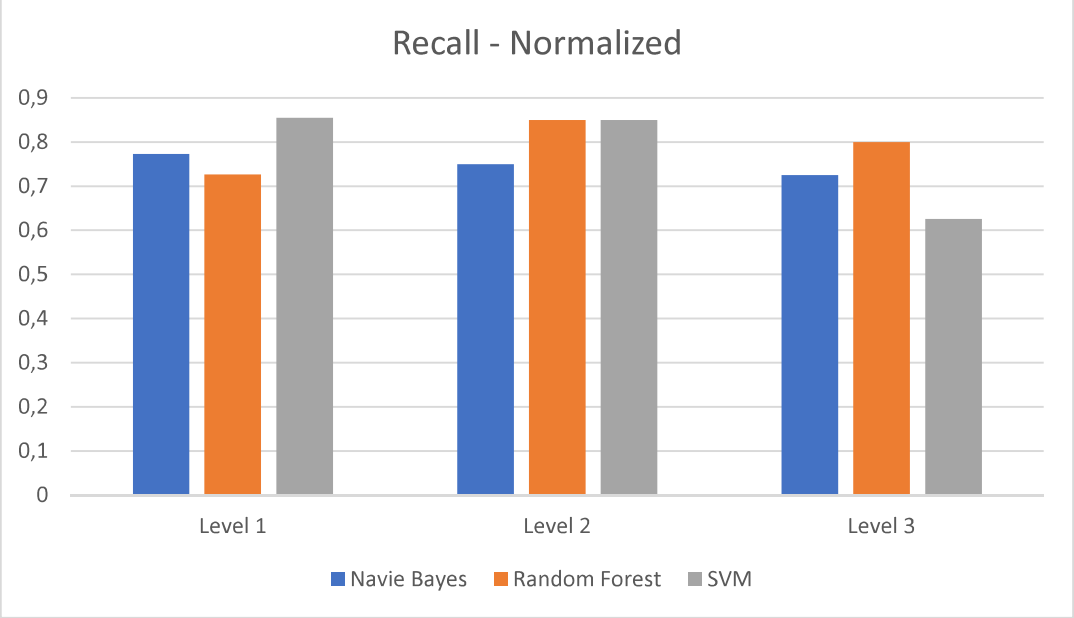

Figure 24: Normalized recall values obtained by traditional machine learning methods for ASR and OCR

Unnormalized recall values obtained by traditional machine learning methods for ASR and OCR can be found in Figure 26.

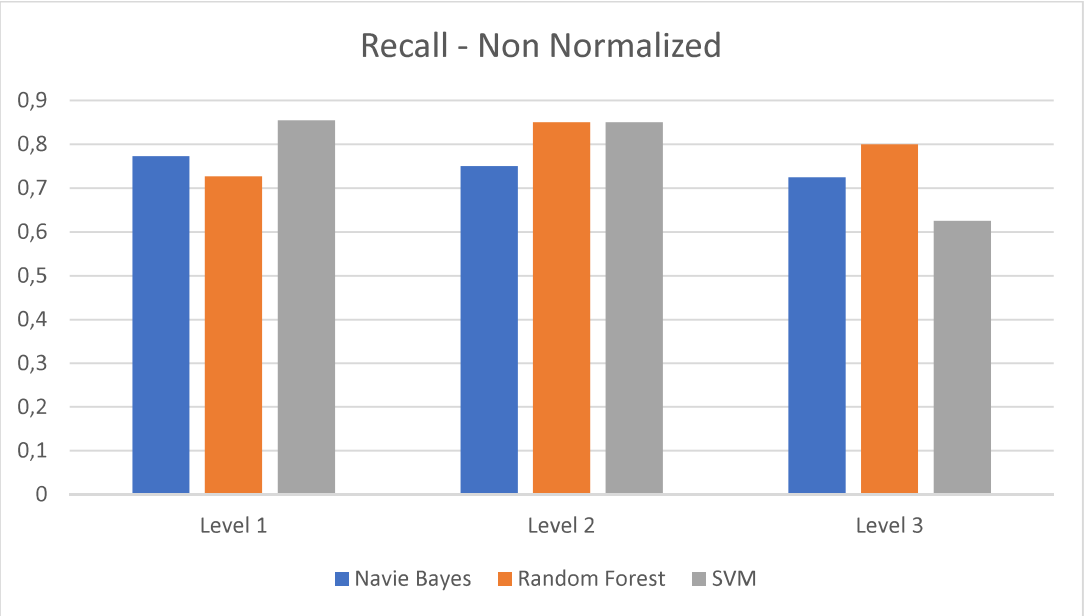

Figure 25: Unnormalized recall values obtained by traditional machine learning methods for ASR and OCR

CHAPTER VIII
DISCUSSION

In this thesis, ASR outputs are extracted from a dataset of 110 videos using ASR. To accomplish this, APIs from the website "https://docs.rev.ai/api/streaming/" are utilized. Using the same dataset as in the thesis of Ağzıyağlı [5], the textual data extracted with OCR and the scenario in which OCR and ASR are used together, are analyzed simultaneously.

After textual data extraction using OCR and ASR, data preprocessing for Naïve Bayes, SVM, and Random Forest classification methods are performed. The data preprocessing operations are discarding unnecessary words by converting the textual data to lowercase, checking the words in the English dictionary and removing words that are incorrectly put by OCR or ASR; determining the word frequency, normalizing the word frequencies and non-normalizing the word frequencies. The texts are then converted to the ".arff" format and classified using the weka program.

In this section, comparisons between traditional classification algorithms are made. In this section, research is conducted to determine which classification algorithm should be used for which indexing techniques. Only normalized groups are used for comparisons in OCR's analysis of Ağzıyağlı's [5] thesis, as all data are normalized according to the frequency plane.

8.1 SVM RESULTS

In this section, it is going to be discussed the outputs of ASR, OCR, and the method that uses both ASR and OCR. Only the results found in accordance with the 90% similarity rate in the OCR-generated thesis [5] are compared.

SVM classification is used to classify Level 1 datasets using ASR and OCR. Level 1 datasets contain the most video files overall. Level 1 analysis has also examined ASR, OCR, and both ASR and OCR in conjunction with SVM. As seen in the graphs in the Results section, SVM also achieves excellent results with the other three methods. Although it yields very good results with these three techniques, it is

evident that ASR analysis yields the most accurate results. A combination of ASR and OCR is observed to yield a less accurate result.

Level 2 datasets are classified using ASR, OCR, and both ASR and OCR, as well as the SVM classification algorithm. Level 2 datasets have the least number of distinct classes. While Level 1 and Level 3 datasets contain four classes, Level 2 datasets contain only three. In this analysis, the highest accuracy rate is obtained only when the ASR method is used. The lowest accuracy rate is obtained only when the OCR method is used.

Level 3 datasets are classified according to the ASR, OCR and both ASR and OCR methods and are classified by the SVM classification method. The least number of video files is present in Level 3 datasets. Using only ASR in the SVM classification analysis of Level 3 datasets has yielded the highest accuracy. The scenario in which ASR and OCR are utilized concurrently has yielded the lowest accuracy.

According to the obtained results, the SVM classification method has the highest accuracy rates in the scenario in which only the ASR method is used. When results are not normalized, accuracy rates have been observed to decrease.

Detailed information about the accuracy result using the SVM algorithm can be found in Table 3.

Table 3: Accuracy results using Support Vector machine algorithm

Accuracy Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	%91,81	%91,25	%97,5
OCR	%88	%73	%67
ASR + OCR	%85,4545	%85	%62,5

In the analysis of the Level 1 data set, only the ASR method has yielded the highest value of precision. Even though the scenario where ASR and OCR methods are used together has the lowest precision value, it is still quite efficient with a value of 0.874%.

In the analysis of the Level 2 dataset, the highest precision value is observed when only ASR is used, whereas the lowest precision value is observed when only OCR is used.

In the analysis of the Level 3 dataset, the highest precision value is observed when only ASR is used, while the lowest precision value is observed when both ASR and OCR are used.

When the precision rates are examined, ASR has yielded positive results at all three levels. Using only OCR has resulted in the lowest value for the least class in the level 2 dataset. In the scenario where both ASR and OCR are utilized, it has provided the lowest value at least at the 3rd data set level.

Detailed information about the precision result using the SVM algorithm can be found in Table 4.

Table 4: Precision results using Support Vector machine algorithm

Precision Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	0,921	0,919	0,977
OCR	0,883	0,731	0,827
ASR + OCR	0,874	0,853	0,677

Only the ASR method has produced the highest recall values across Level 1, Level 2, and Level 3 analyses. The lowest recall values are observed at the 3rd level of the method, when only OCR and both ASR and OCR are used, which corresponds to the lowest level of the data set.

In the analysis that utilized only OCR, it is also observed that the Recall value for the level 2 dataset, which contains fewer classes, is partially low.

Detailed information about the recall result using the SVM algorithm can be found in Table 5.

Table 5: Recall results using Support Vector machine algorithm

Recall Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	0,918	0,913	0,975
OCR	0,873	0,725	0,675
ASR + OCR	0,855	0,85	0,625

The highest F-score is only observed when ASR is implemented. Again, in the case where only ASR is used, the F-score is quite high, particularly for level 3, where the data set is the least. In the method employing only OCR, the lowest F-score is observed for level 3 with the least data points. In the study employing OCR, the value of the F-score has decreased as the dataset size decreased. In contrast, when only ASR is utilized, the F-score value has increased as the dataset size decreases.

Detailed information about the F1 score result using the SVM algorithm can be found in Table 6.

Table 6: F1 score results using Support Vector machine algorithm

F1 Score Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	0,912	0,912	0,975
OCR	0,865	0,724	0,647
ASR + OCR	0,844	0,85	0,61

8.2 NAÏVE BAYES RESULTS

In this section, it is going to be discussed the outputs of ASR, OCR, and the method that uses both ASR and OCR. Only the results found in accordance with the 90% similarity rate in the OCR-generated Ağzıyağlı's thesis [5] are compared.

Using ASR and OCR, Level 1 datasets are classified with Naïve Bayes classification. Level 1 datasets contain the greatest total number of video files. ASR, OCR, and ASR and OCR in conjunction with Naïve Bayes are also examined at the level 1 analysis. In general, as evidenced by the graphs in the section on results, Naïve Bayes only has produced good results for unnormalized ASR and OCR results.

On Level 1, Level 2, and Level 3 datasets, ASR, OCR, and both ASR and OCR are used to evaluate the Naïve Bayes algorithm. It can be said that, among these three levels, the OCR experiment has produced the most consistent results.

Detailed information about the accuracy result using the Naïve Bayes algorithm can be found in Table 7.

Table 7: Accuracy results using Naïve Bayes machine algorithm

Accuracy Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	%32,7273	%41,25	%60
OCR	%89	%77	%75
ASR + OCR	%77,2727	%75	%72,5

These three levels of data are utilized for precision value analysis. In support of the accuracy values in this experiment, OCR analysis has yielded the highest precision values. From the ASR experiment, it is impossible to obtain values. The OCR indexing method has yielded the highest precision result value during level 1 analysis.

Precision, on the other hand, indicates how many of the estimated positive values are in fact positive. In this context, it is typical for highly precise values to have high precision rates.

Detailed information about the precision result using the Naïve Bayes algorithm can be found in Table 8.

Table 8: Precision results using Naïve Bayes machine algorithm

Precision Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	-	-	-
OCR	0,922	0,792	0,827
ASR + OCR	0,811	0,757	0,742

In Level 1, Level 2, and Level 3 analyses, it is determined that the Recall values for all three methods are average. It has been observed that ASR Recall results are higher compared to others when examined in detail.

Detailed information about the recall result using the Naïve Bayes algorithm can be found in Table 9.

Table 9: Recall results using Naïve Bayes machine algorithm

Recall Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	0,973	0,943	0,975
OCR	0,882	0,775	0,75
ASR + OCR	0,773	0,75	0,725

The highest F1 score is observed only when OCR is applied. Significant results cannot be achieved when only ASR is used. In the method using only OCR, the highest F1 score is obtained in the level 1 cluster, where the data set is large. In studies using OCR and both OCR and ASR, the value of the F1 score has decreased as the dataset size decreased.

Detailed information about the F1 score result using the Naïve Bayes algorithm can be found in Table 10.

Table 10: F1 score results using Naïve Bayes machine algorithm

F1 Score Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	-	-	-
OCR	0,89	0,777	0,761
ASR + OCR	0,786	0,739	0,726

8.3 RANDOM FOREST RESULTS

In this section, it is discussed the outputs of ASR, OCR, and the method that uses both ASR and OCR. Only the results found in accordance with the 90% similarity rate in the OCR-generated Ağzıyağlı's thesis [5] are compared.

In the analysis of Level 1, all indexing methods have yielded similar results. For the level 2 and level 3 analyses, it is obtained almost better results using only ASR, despite all the positive outcomes. In general, the scenario where the smallest dataset is used at level 3 and only the ASR has yielded the best result.

Detailed information about the accuracy result using the Random Forest algorithm can be found in Table 11.

Table 11: Accuracy results using Random Forest machine algorithm

Accuracy Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	%72,7273	%86,25	%87,5
OCR	%75	%79	%70
ASR + OCR	%72,7273	%82,5	%82,5

In Level 1, Level 2, and Level 3 analyses, cannot be gained any results for Precision Values. The highest precision value is gained by the ASR method.

Detailed information about the precision result using the Random Forest algorithm can be found in Table 12.

Table 12: Precision results using Random Forest machine algorithm

Precision Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	-	0,874	0,904
OCR	-	0,783	0,782
ASR + OCR	-	0,89	0,869

Only the ASR method has produced the highest recall values across in Level 3 analysis. The lowest recall value is observed at the 2nd level of the method, when only OCR, which has the least number of distinct classes.

In Level 1 analysis, the same recall results were obtained in all indexing methods002E

Detailed information about the recall result using the Random Forest algorithm can be found in Table 13.

Table 13: Recall results using Random Forest machine algorithm

Recall Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	0,727	0,863	0,875
OCR	0,727	0,688	0,70
ASR + OCR	0,727	0,85	0,80

Only the ASR method has produced the highest F1 values across in Level 3 analysis. The lowest F1 value is observed at the 2nd level of the method, when only OCR, which has the least number of distinct classes.

Detailed information about the F1 score result using the Random Forest algorithm can be found in Table 14.

Table 14: F1 score results using Random Forest machine algorithm

F1 Score Values			
Indexing Methods	**Level 1**	**Level 2**	**Level 3**
ASR	-	0,861	0,876
OCR	-	0,64	0,71
ASR + OCR	-	0,852	0,799

CHAPTER IX
CONCLUSION AND RECOMMENDATIONS

In this chapter, the concluding remarks on the thesis are once again presented, along with a summary and conclusion of the thesis project. Both the contribution of the thesis to the field and its limitations are evaluated. In addition, based on the experiments conducted and results obtained, recommendations are made regarding aspects that can be investigated in future studies.

The recent COVID-19 pandemic has demonstrated the viability and efficacy of distance learning. Due to this circumstance, the number of lecture videos on the WWW increases daily. This effect indicates that online education will rapidly expand in the not-too-distant future. Due to this tendency, it has become more difficult to locate the desired lecture videos in the repositories. In this thesis, classification is performed using three traditional machine learning algorithms and three indexing techniques. In this context, the outcomes are also investigated.

In this thesis, an indexing technique and a classification algorithm that should be used along with it, are proposed in order to find lecture videos from large repositories with the highest accuracy. These suggestions are made to facilitate distance education and to match the right person together with the right educational content.

In this study, there are three different levels of datasets. Using OCR and ASR, textual data are extracted from these three-level datasets. The extracted data are classified using three different classification algorithms. The accuracy rates of algorithms utilized in datasets vary depending on the datasets.

The SVM classification algorithm has had the highest success rate in terms of accuracy, precision, recall, and F1 score when applied to the text information extracted via ASR.

Level 1 and Level 2 datasets are less balanced than the entire dataset. SVM methods have provided the highest accuracy, recall, precision, and F1 score ratios for these two levels.

When the ASR-extracted text information is classified by the Naïve Bayes algorithm, Precision values are incalculable. It has been determined that the most effective indexing method for all three levels of the Naïve Bayes algorithm is an analysis using both OCR and ASR. It has been seen that the most inefficient indexing method of the Naïve Bayes algorithm is ASR.

The SVM method is the most effective traditional machine learning method for evaluating the text information extracted by the ASR method by calculating the arithmetic mean of the results obtained in terms of accuracy, recall, precision, and F1 score ratios.

Although very productive results were obtained in the analysis using ASR and OCR together, it is not as good as the analysis using ASR alone. The main reason for this is that the OCR analysis results in a lot of meaningless data. Today, ASR technologies work with almost zero errors.

Although OCR is very successful in printed texts when processing image frames, unfortunately, it does not work well in handwritten texts. In addition to that, the image frames in the lecture videos may be insufficient to give an idea about the video. In these circumstances, it is logical that ASR has provided a more sophisticated result than OCR. In order to accomplish ASR, APIs from the website "https://docs.rev.ai/api/streaming/" are utilized. Imaginably enhanced results can be obtained if Amazon's transcribe service is used in prospective studies.

All the lecture videos used in this research, except video 105, have a speech. However, occasionally, lecture videos can be quite diverse in some repositories. It may not be possible to extract textual information for ASR. In order to overcome such a situation, it is necessary to make an analysis again with a dataset consisting of lecture videos with a lack of explanatory audio content.

In this thesis, ASR and OCR are compared with accuracy, precision, recall and F1 score metrics. Although the comparison of these metrics has provided a crucial point of view to the matter, in order for these metrics to be used in the education and software industry (edtech), performance metrics play as much of a serious role. Performance comparison is just as important for the productization of these technologies in the education and software industry. In this context, performance comparisons of ASR and OCR techniques should also be made in the following studies.

Three different traditional classification algorithms are used in the analysis in this thesis. As a suggestion regarding subsequent studies and research, classification can be made with K-Nearest Neighbors classification algorithm or convolutional neural network (CNN) algorithms.

This thesis should be viewed as a modest contribution and the initial step towards a very important topic: content-based search and retrieval. This is a tremendous research field, and it is expected that in the near future a great deal of research will be conducted.

REFERENCES

[1] Cross Jay (2004), "An informal history of eLearning", *On the Horizon*, Vol 12, No. 3, pp. 103-110.

[2] Fisler Joël and Schneider Franziska (2009),"Creating, handling and implementing e-learning courses using the Open source tools OLAT and eLML at the University of Zurich", Proceedings *of the world congress on engineering and computer science,* Vol. 1, pp 1-8.

[3] Taivalsaari Antero, Mikkonen Tommi and Systä Kari (2014), "Liquid software manifesto: The era of multiple device ownership and its implications for software architecture", *2014 IEEE 38th Annual Computer Software and Applications Conference,* pp. 338-343, Vasteras.

[4] McCue TJ (2018), *E-learning climbing to $325 Billion by 2025 UF Canvas Absorb Schoology Moodle*, https://www.forbes.com/sites/tjmccue/2018/07/31/e-learning-climbing-to-325-billion-by-2025-uf-canvas-absorb-schoology-moodle/?sh=308b4b263b39, DoA. 27.02.2023.

[5] Ağzıyağlı Veysel Sercan and Oğul Hasan (2020), *Ders videolarının içerik tabanlı erişimi* (Master Thesis), Başkent Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

[6] Rish Irina (2001), "An empirical study of the naive Bayes classifier", *workshop on empirical methods in artificial intelligence*, Vol. 3, No. 22, pp. 41-46.

[7] Biau Gerard and Scornet Erwan (2016), "A random forest guided tour", *Test*, No. 2, Vol 25, pp. 197-227.

[8] Pavlidis Paul, Wapinski Ilan, and Noble William Stafford (2004), "Support vector machine classification on the web", *Bioinformatics*, Vol. 20, No. 4, pp. 586-587.

[9] Chand Dipesh and Ogul Hasan (2020), "Content-Based Search in Lecture Video: A Systematic Literature Review", *In 2020 3rd International Conference on Information and Computer Technologies (ICICT)*, pp. 169–176, San Jose.

[10] Yang Haojin and Meinel Christoph (2014), “Content based lecture video retrieval using speech and video text information”, *IEEE transactions on learning technologies*, Vol.*7,* No. 2, pp. 142-154.

[11] Salton Gerard, Wong Anita and Yang Chung-Shu (1975), “A vector space model for automatic indexing”, *Communications of the ACM*, Vol. 18, No. 11, pp. 613-620.

[12] Salton Gerard and Buckley Christopher (1988), “Term-weighting approaches in automatic text retrieval”, *Information processing and management*, Vol. *24,* No.5, pp. 513-523.

[13] Haubold Alexander and Kender John (2005), “Augmented segmentation and visualization for presentation videos”, *In Proceedings of the 13th annual ACM international conference on Multimedia,* pp. 51-60, New York.

[14] Chen Scott and Gopalakrishnan Ponani (1998), “Speaker, environment and channel change detection and clustering via the bayesian information criterion” *In Proc. DARPA broadcast news transcription and understanding workshop,* Vol. 8, pp. 127-132.

[15] Adcock John, Cooper Matthew, Denoue Laurent, Pirsiavash Hamed and Rowe Lawrence (2010), “Talkminer: a lecture webcast search engine”, *In Proceedings of the 18th ACM international conference on Multimedia*, pp. 241-250, Vancouver.

[16] Chivadshetti Pradeep, Sadafale Kishor and Thakare Kalpana (2015), “Content based video retrieval using integrated feature extraction and personalization of results”, *In 2015 International Conference on Information Processing (ICIP),* pp. 170-175, Pune.

[17] Alharbi Sadeen, Alrazgan Muna, Alrashed Alanoud, AlNomasi Turkiayh, Almojel Raghad, Alharbi Rimah, Saja Alturki, Sahar Alshehri and Almojil Fatimah (2021), “Automatic speech recognition: Systematic literature review”, *IEEE Access*, Vol. 9, pp. 131858-131876

[18] Sateli Bahar, Cook Gina and Witte Ren (2013), “Smarter mobile apps through integrated natural language processing services”, *In International Conference on Mobile Web and Information Systems,* pp. 187-202, Montréal.

[19] Tate Laura(2023), *The Difference Between Speech and Voice Recognition* [Image], Voice User Interface Technology, Retrieved from https://www.kardome.com/blog-posts/difference-speech-and-voice-recognition, DoA. 27.02.2023.

[20] Mithe Ravina, Indalkar Supriya and Divekar Nilam (2013), "Optical character recognition", *International journal of recent technology and engineering*, Vol. 2, No. 1, pp. 72-75.

[21] Zelic Filip and Sable Anuj (2023), *A comprehensive guide to OCR with Tesseract, OpenCV and Python*, *https://nanonets. com/blog/ocr-withtesseract*, DoA. 27.02.2023.

[22] Naushad Raoof (2020), *OCR-Tesseract with Image Pre-processing* [Image], Medium, Retrieved from https://medium.com/swlh/ocr-tesseract-with-image-pre-processing-fb415d3be4ee, DoA. 27.02.2023.

[23] Debnath Lokenath and Basu Kanadpriya (2015), "A short history of probability theory and its applications", *International Journal of Mathematical Education in Science and Technology*, Vol. 46, No. 1, pp. 13-39.

[24] Alam Bashir(2022), *Naive-Bayes-Classification*[Image], Retrieved from https://hands-on.cloud/naive-bayes-classifier-python-tutorial/#h-what-is-naive-bayes-classification, DoA. 27.02.2023.

[25] Leung Ming (2007), "Naive bayesian classifier", *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, Vol. 2007, pp. 123-156.

[26] Jadhav Sayali and Channe HP (2016), "Comparative study of K-NN, naive Bayes and decision tree classification techniques", *International Journal of Science and Research (IJSR)*, Vol. 5, No. 1, pp. 1842-1845.

[27] Jakkula Vikramaditya (2006), "Tutorial on support vector machine (svm)", *Washington State University*, Vol. 37, No. 2.5, p. 3.

[28] Arora Avi(2021), *8 Unique Machine Learning Interview Questions about Random Forests - Analytics Arora* [Image], Analytics Arora, Retrieved from https://analyticsarora.com/8-unique-machine-learning-interview-questions-about-random-forests, DoA. 27.02.2023.

[29] IBM(2023), *What is Random Forest?,* https://www.ibm.com/topics/random-forest, DoA. 27.02.2023.

[30] Aher Sunita and Lobo (2011), “Data mining in educational system using weka”, *International Conference on Emerging Technology Trends (ICETT),* Vol. 3, pp. 20-25, Solapur.