ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

A VIDEO DATASET OF INCIDENTS & VIDEO-BASED INCIDENT CLASSIFICATION

M.Sc. THESIS

Duygu SESVER

Department of Computer Engineering

Computer Engineering Programme

JUNE 2022

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

A VIDEO DATASET OF INCIDENTS & VIDEO-BASED INCIDENT CLASSIFICATION

M.Sc. THESIS

Duygu SESVER
(504191509)

Department of Computer Engineering

Computer Engineering Programme

Thesis Advisor: Prof. Dr. Hazım Kemal EKENEL

JUNE 2022

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

FELAKET VİDEO VERİSETİ & VİDEO-TABANLI FELAKET SINIFLANDIRMASI

YÜKSEK LİSANS TEZİ

Duygu SESVER
(504191509)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Hazım Kemal EKENEL

HAZİRAN 2022

Duygu SESVER, a M.Sc. student of ITU Graduate School student ID 504191509 successfully defended the thesis entitled “A VIDEO DATASET OF INCIDENTS & VIDEO-BASED INCIDENT CLASSIFICATION”, which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Hazım Kemal EKENEL** Istanbul Technical University	....
Jury Members :	**Prof. Dr. Mine Elif KARSLIGİL YAVUZ** Yıldız Technical University	....
	Prof. Dr. Behçet Uğur TÖREYİN Istanbul Technical University	....

Date of Submission : **02 June 2022**
Date of Defense : **22 June 2022**

To my family and friends,

FOREWORD

I would like to thank my mother Nurhayat SESVER and my father Melih SESVER for their infinite love and their faith in me. I would not be able to make it without you.

I would also like to thank

- My advisor Prof. Dr. Hazım Kemal EKENEL for being an excellent advisor, for his valuable comments, and for creating new opportunities for me.
- My friend Çağrı Emre YILDIZ for always supporting me, finding a solution every time I get into a dead-end, and making work fun.
- My teammates in this project (Alp Eren GENÇOĞLU, Zehra GUNINDI, Faeze HABIBI, Ziya Ata YAZICI) for making this possible and for their passion.
- To all my friends who helped me in the data collection part and whose support I always felt by my side.
- The Scientific and Technological Research Council of Turkey (TUBITAK) ARDEB 1001 Grant No 121E408 for funding this study.

June 2022

Duygu SESVER
Computer Engineer

TABLE OF CONTENTS

Page

ABBREVIATIONS

BCEW	**:** Binary Cross-Entropy with Logits Loss
BERT	**:** Bidirectional Encoder Representations from Transformers
BLEU	**:** BiLingual Evaluation Understudy
CELoss	**:** Cross-Entropy Loss
ID	**:** Identification
LR	**:** Learning Rate
LSTM	**:** Long-Short Term Memory
MLP	**:** Multi-Layer Perceptron
NLP	**:** Natural Language Processing
SVM	**:** Support Vector Machine
USA	**:** United States of America
USSR	**:** Union of Soviet Socialist Republics
VGG	**:** Visual Geometry Group
VIDI	**:** Video Dataset of Incidents
ViT	**:** Vision Transformer
VTAB	**:** Visual Task Adaptation Benchmark
WMT	**:** Workshop on Statistical Machine Translation

LIST OF TABLES

LIST OF FIGURES

A VIDEO DATASET OF INCIDENTS & VIDEO-BASED INCIDENT CLASSIFICATION

SUMMARY

Nowadays, the occurrence of natural disasters, such as fires, earthquakes, and floodings have increased in our world. Detection of incidents and natural disasters became more important when action is needed. Social media is one of the data sources to see natural disasters and incidents thoroughly and immediately. There are a lot of studies to detect incidents in the literature. However, most of them are using still images and text datasets. The number of video-based datasets is limited in the literature. Moreover, existing video-based studies have a limited number of class labels. Motivated by the lack of publicly available video-based incident datasets, a diverse dataset with a high number of classes is collected which we named as Video Dataset of Incident (VIDI).

The collected dataset has 43 classes which are exactly the same as the ones in the previously published Incidents Dataset. It includes both positive and negative samples for each class as in the Incidents Dataset. The dataset contains 8.881 videos in total, 4.534 of them are positive samples and 4.347 of them are negative samples. Approximately, there are 100 videos for each positive and negative class. Video duration is ten seconds on average. YouTube is used as the source of the videos and video clips are collected manually.

Positive examples of the dataset consist of natural disasters such as landslides, earthquakes, floods, vehicle accidents, such as truck accidents, motorcycle accidents, and car accidents, and the consequences of these events such as burned, damaged, and collapsed. Positive samples may contain multiple labels per video and image. It means video can belong to more than one class category. On the other hand, negative samples do not contain the disaster of that class. Samples in the negative can be instances that the model can easily confuse. For instance, the negative example of "car accident" class can be a normal car driving or a video that includes a "flooded" incident, since it contains "flooded" incident but it does not include a "car accident".

It is aimed to ensure diversity in the dataset while collecting videos. Videos from different locations for each positive and negative sample are collected to increase diversity. Another approach to increase diversity is to look for videos in various languages to capture the styles of different cultures and include region-specific events. Six languages are used and they are as follows: Turkish, English, Standard Arabic, French, Spanish, and Simplified Chinese. When these languages are not sufficient, videos are queried in different languages, too.

After the dataset is collected, various experiments are performed on it. While performing these experiments, the latest video and image classification architectures are used on both existing image-based and newly created video-based incident datasets.

The first part of the study is conducted by using only positive samples and negative samples are included in the second part.

One of the motivations was to explore the benefit of using video data instead of images for incident classification. To investigate this need, the Vision Transformer and TimeSformer architectures are trained using only positive samples of both datasets for the classification of the incident. Top-1 and top-5 accuracies are used as evaluation metrics.

ViT which is designed for the images is performed on the Incidents Dataset. Then, TimeSformer which is designed for multi-frame data is performed on the collected video-based dataset. Eight frames are sampled from each video in the collected dataset and used for the TimeSformer multi-frame training. Since datasets and architectures are not the same in these experiments, it would not be a fair comparison between the image and the video datasets. So, the TimeSformer architecture is performed also on the Incidents Dataset and the ViT is performed also on the VIDI. Input data is adapted according to the requirements of the architectures. In the video classification architecture, each image from the Incidents Dataset was treated as a single-frame video. In the image classification architecture, the middle frame of the input video is used as an image. Finally, to be able to show the impact of using multiple-frames in the incident classification, the TimeSformer architecture is performed also with a single-frame from a video dataset. The same downsampling method is applied and the middle frame is used for the training.

TimeSformer achieved 76.56% accuracy in the multi-frame experiment, while 67.37% accuracy is achieved in the single-frame experiment on the collected dataset. This indicates that using video information when available improves incident classification performance.

In the experiments, the performance of the state-of-the-art, ViT and TimeSformer architectures, for incident classification is also evaluated. The used method in the Incidents Dataset paper and state-of-the-art image classification architecture performance are compared. The used approach in the Incidents Dataset paper was using ResNet-18 architecture. ViT and TimeSformer achieved higher accuracies than the ResNet-18 on the image-based Incidents Dataset. While the Resnet-18-based model achieved 77.3% accuracy, ViT and TimeSformer achieved 78.5% and 81.47% top-1 accuracy, respectively.

Additionally, the performance of ViT and TimeSformer is compared using both datasets in their single-frame version. TimeSformer achieved 67.37% and ViT achieved 61.78% on the single-frame version of the video-based dataset. Moreover, TimeSformer performed 81.47% and ViT performed 78.5% on the image dataset. TimeSformer is found to be superior to ViT on both datasets. However, the results in the collected dataset are lower than those obtained in the Incidents Dataset. There could be two main reasons for this: (1) the image-based dataset contains more examples for training, so systems can learn better models. (2) The collected dataset contains more difficult examples for classification.

The second part of the study includes negative samples. By using both positive and negative samples, binary classification models are trained for all classes. The main idea was to measure the performance that a model could detect whether or not that disaster occurs in a given video. Therefore, 43 separate models are trained. As a result, the best accuracy is achieved by the "landslide", and "dirty contamined" classes. The model got the lowest accuracy in the detection of "blocked" disasters.

Finally, one more classification experiment has been run on VIDI. This experiment uses negative samples as the 44th class. For the 44th class, 100 videos that do not include any incidents are selected from the negative samples. By using these classes, 72.18% accuracy is achieved for this experiment.

In summary, a highly diverse disaster dataset with many classes is presented in this study. For the classification tasks, the performances of the recent video and image classification architectures on video and image datasets are compared, and binary classification experiments are done for each class.

FELAKET VİDEO VERİSETİ & VİDEO-TABANLI FELAKET SINIFLANDIRMASI

ÖZET

Son günlerde dünyada, doğal afetlerin ve felaketlerin sayısında artış görülmeye başlandı. Bu gibi durumlar meydana geldiğinde hızlı aksiyon almak önemli bir hal aldı. Sosyal medya bu konuda önemli bir veri kaynağıdır. Dünyanın herhangi bir yerinde felaket meydana geldiğinde, bu olaylara ait görüntüler, sosyal medyada hemen yer alabilir. Bu sebeple, bu gibi durumlarda sosyal medyayı verimli kullanmak gerekmektedir.

Literatürde, felaketlerin tespiti için pek çok çalışma yapılmıştır. Ancak, var olan veri setleri, çoğunlukla metin ve resim üzerinedir ve literatürdeki video tabanlı felaket veri setleri sınırlı sayıdadır. Var olan veri setleri ise, sınırlı sayıda felaket sınıf etiketine sahiptir. Bu eksiklikten yola çıkarak, bu çalışma ile literatüre çeşitliliği fazla ve yüksek sayıda sınıfa sahip VIDI adında video tabanlı bir veri seti sunulması hedeflenmiştir.

Yüksek sınıf sayısına sahip olup çeşitliliği yüksek olan bir resim tabanlı veri seti, halihazırda literatürde mevcuttur. Bu çalışmada, resim veri setini tamamlayacak, video tabanlı bir veri seti oluşturulması hedeflenmiştir. Oluşturulan veri setinin sınıfları, var olan felaket resim veri setinin sınıfları ile aynı olup, 43 adet farklı sınıf sayısına sahiptir. Oluşturulan veri seti, her bir felaket sınıfı için pozitif ve negatif veriler içermektedir.

Veri seti, toplamda 8.881 adet videodan oluşmaktadır. Bunlardan 4.534 tanesi pozitif örnekler olup felaket içermekteyken, 4.347 tanesi 43 sınıfın negatif örnekler olarak veri setinde bulunmaktadır. Her bir sınıf için yaklaşık 100 tane pozitif ve negatif video bulunmaktadır. Video kliplerinin süresi minimum 2 saniye, maksimum 60 saniyedir. Çoğu video, ortalama on saniye sürmektedir. Veri setindeki bütün videoların kaynağı olarak YouTube kullanılmıştır.

Veri setinin pozitif örnekleri deprem, sel, heyelan gibi doğal afetler, araba kazası, kamyon kazası, motosiklet kazası gibi araç kazaları ve bu olayların sonuçlarını içeren hasarlı, yanmış, yıkılmış durumlarını içeren videolardan oluşur. Pozitif örnekler, birden fazla etiket ile etiketlenmiş olabilir. Bunun sebebi, bir videonun birden fazla felaketi içerebilmesidir. Örneğin, yanmış bir ev hem yanmış hem hasarlı etiketiyle etiketlenebilir veya bir araba kazası hem trafik kazası hem de bloklu olarak etiketlenebilir.

Negatif örnekler ise, her sınıf için o felaketi içermeyen videolardan oluşur. Negatif örnekler toplanırken, bu verilerin, pozitif örneklerle kolayca karıştırılabilecek videolar olmasına özen gösterilmiştir. Örneğin, "araba kazası" sınıfının negatifi olarak, kaza olmayan araba sürüş videoları ya da "sel" içeren videolar, araba kazası içermediği için, araba kazasının negatifi olarak veri setine eklenmiştir.

Veriler toplanırken, çeşitliliğin yüksek olması hedeflenmiştir. Bu sebeple, her bir sınıfın pozitif ve negatif örnekleri için farklı lokasyonlardan videolar toplanmıştır. Çeşitliliği artırmaya yönelik bir diğer yaklaşım da, farklı kültürlerin tarzlarını elde etmek ve o bölgeye özgü olayları dahil etmek için çeşitli dillerde videolar aramaktır. Temel olarak şu altı dil kullanılmıştır: Türkçe, İngilizce, Standart Arapça, Fransızca, İspanyolca, Basitleştirilmiş Çince. Bu dillerde toplanan video sayısı yeterli olmadığında, daha fazla video toplamak amacıyla, veri setinde Hintçeye, Almancaya ve birkaç farklı dile de yer verilmiştir.

Bu dillerin seçiminde, dünya genelinde konuşulma sıklığı ve bu dillerin konuşulduğu bölgelerdeki çevre şartları dikkate alınmıştır. Örneğin, farklı bölge veya ülkelerdeki binaların görünümü kültür, konum ve malzeme mevcudiyetine bağlı olarak birbirinden farklı olabilir. Amerika'da yapı malzemesi olarak yaygın olarak ahşap kullanılırken, Orta Doğu'da kireçtaşı daha yaygındır. Ayrıca, farklı iklim olaylarının önlemleri ve sonuçları, beklendiği ülkelerde ve beklenmediği ülkelerde oldukça farklıdır.

Bazı afetlerin, belirli bölgelerde yaşanması nedeniyle, her sınıfta bu dillerin bulunması mümkün olmayabilir. Örneğin, derecho -Türkçeye, düz ve sert esen rüzgar olarak çevrilebilir ancak Türkiye'de görülmemektedir.-, çoğunlukla Kuzey Amerika'da var olan bir olaydır. Bu nedenle, derecho video klipleri ağırlıklı olarak İngilizce olarak toplanmıştır. Bu etiketteki problemi düzeltmek için bu felaketin yaşandığı ülkelerin listesi bulunmuş ve bu bölgelere göre farklı dillerde aramalar yapılmıştır.

Veri toplama sırasında karşılaşılan bir diğer engel ise anahtar kelimelerin diller arası çevirisi sırasında ortaya çıkan uyuşmazlıklar olmuştur. Bazı diller için yapay sinir ağları tabanlı çeviri uygulamaları kullanılmıştır. Kullanılan çeviri uygulamasının çalışma prensipleri gereği, hedef dilde çevirilmek istenen olaya ait bir kelime olmadığında, o olaya en yakın isim veya isim tamlaması önerilmektedir. Tropikal siklon bu duruma örnek olarak verilebilir. "Tropikal siklon" etiketi Arapçaya çevrildiğinde, tercüme edilen kelimenin kasırga ile aynı anlama geldiği görülmektedir. Bu sorunlar, hedef dili anadili olarak kullanan insanlar ile birebir etkileşime geçilerek çözülmeye çalışılmıştır.

Çeviri sorunlarının yanı sıra, bir diğer problem de nükleer patlama gibi bazı etiketler için video bulunmasıdır. Nükleer patlamalar için mevcut videolar 1945 ile 1963 yılları arasında Amerika Birlesik Devletleri ve Sovyet Sosyalist Cumhuriyetler Birliği'nde kaydedilmiştir. O yıllardan sonra atmosferik testler yasaklandığından dolayı, bu etikete sahip sınırlı sayıda video vardır. Bu nedenle, toplanan videolar çoğunlukla aynı sahnenin, farklı kanallardan kopyalarını içermektedir.

Veri seti toplandıktan sonra, üzerinde çeşitli deneyler yapılmıştır. Bu deneyleri gerçekleştirirken, hem var olan resim tabanlı hem yeni oluşturulan video tabanlı felaket veri setlerinde en yeni video ve resim sınıflandırma mimarileri kullanılmıştır. Adil bir performans karşılaştırması yapılmasına özen gösterildiği için, her iki veri seti de aynı sınıf etiketlerine sahiptir. Çalışmanın ilk bölümünde sadece pozitif veriler kullanılırken, ikinci bölümünde ise hem pozitif hem de negatif veriler kullanılmıştır.

Bu çalışmanın motivasyonlarından biri, felaket sınıflandırması için resim yerine video verileri kullanmanın faydasını gözlemlemektir. Bunu sağlamak için ilk aşamada, Vision Transformer ve TimeSformer mimarileri, felaket sınıflandırması için her iki veri kümesinin yalnızca pozitif örnekleri kullanılarak eğitilmiştir.

Kullanılan mimarilerden bahsetmek gerekirse, Vision Transformer mimarisi, resim sınıflandırma amaçlı tasarlanmıştır. Temel olarak doğal dil işleme görevleri için kullanılan ve herhangi bir konvolüsyonel katman kullanmayan Transformer bloklarından oluşur. Transformer bloklarının birleştirilmesi ile ViT mimarisi oluşturularak resimlere uygulanmıştır ve resim sınıflandırma için literatürde elde edilen doğruluk oranlarından daha iyi bir sonuç elde edilmiştir.

ViT mimarisi çalışma düzeninde, ilk olarak, resimler kare parçalara bölünür ve her biri bir simge olarak ele alınır. Bir sonraki adımda parçalar düzleştirilir ve alt boyuta indirgenir. Parçaların sırası model tarafından bilinmediğinden dolayı, düzleştirilen parçaya konum bilgisi eklenir. Ardından, standart Transformer kodlayıcı blokları, hem parçalarla hem de düzleştirilmiş yamaların konumları ile beslenir. Pek çok veri setinde başarılı sonuçlar elde etmiştir. Bu sebeple, bu çalışmada da sınıflandırma problemleri için bu mimari tercih edilmiştir.

TimeSformer modeli, Vision Transformer'dan ilham alınan bir mimaridir ve ViT mimarisine zamansal boyutta dikkat mekanizmaları eklenerek video işleme için uygun hale getirilmiştir. Vision Transformer'a benzer şekilde TimeSformer da, video işleme için Transformer mimarisini kullanır.

TimeSformer mimarisinde, bir videodaki her bir ekran parçalara ayrılarak doğrusal olarak düzleştirilir. Ardından, bu düzlemlere parçaların uzamsal ve zaman konumları eklenir. Dikkat mekanizmasından geçirilerek sınıflandırma yapılır. TimeSformer mimarisi çalışmasında, tasarlanan beş farklı dikkat mekanizması ile çeşitli deneyler yapılmıştır. En iyi sonucu elde eden bölünmüş uzay-zaman dikkat mekanizması, bu çalışmada kullanılmıştır.

Bu mimariler iki adet veri seti üzerinde eğitilmiştir. Bunlardan ilki Incidents veri setidir. Bu veri seti resim tabanlı olup, 43 adet farklı kategoriden oluşmaktadır. Veri seti, makalenin yazarları tarafından, eğitim, doğrulama ve test kümelerine bölünerek paylaşılmıştır. Eğitim setinde 312.742 örnek bulunurken, doğrulama setinde 17.357 örnek ve test setinde 17.255 örnek bulunmaktadır.

Karşılaştırma yapabilmek amacıyla, VIDI, bu çalışmada oluşturulmuştur ve video tabanlı bir veri setidir. Oluşturulan veri seti, pozitif örnekler ile sınıflandırma deneylerinde kullanılmıştır. Bu deneyler, eğitim için 3.721 video, doğrulama için 408 video ve test için 405 video içermektedir. Toplanan negatif örnekler, pozitif örnekler ile birleştirilerek, her bir sınıf için ikili sınıflandırma deneyleri yapılmıştır. Her sınıfın %10'i doğrulama ve %10'i test işlemleri için kullanılırken, geriye kalan kısmı eğitim için ayrılmıştır.

VIDI kümesi üzerinde 43 sınıflı ve 44 sınıflı felaket sınıflandırma deneyleri yapılmıştır. 43 sınıflı deneylerde VIDI kümesi ile Incidents veri setinin felaket sınıflandırma sonuçları karşılaştırılmıştır. 44 sınıflı deneylerde ise, hiç bir felaket içermeyen 100 adet negatif örnek, veri setine 44. sınıf olarak eklenmiştir. Eklenen 44. sınıf ile tekrardan TimeSformer mimarisi felaket sınıflandırma üzerine eğitilmiş ve negatif örneklerin eklenmesinin sınıflandırma başarısına olan etkisi ölçülmüştür.

43 ve 44 sınıflı deneylerde performansı ölçmek için, en iyi bir ve en iyi beş doğrulukları kullanılmıştır. İkili sınıflandırma deneylerinde, en iyi bir doğruluk metriği kullanılmıştır.

Deney bölümünde, bu mimariler farklı veri setleri ve veri setlerinin farklı versiyonları ile çalıştırılmıştır. Örneğin, resim verilerini işlemek için tasarlanan ViT, resim veri seti üzerinde çalıştırılmıştır. Daha sonra, toplanan video tabanlı veri seti üzerinde, video için tasarlanan TimeSformer çalıştırılmıştır. Toplanan veri kümesindeki her videodan sekiz ardışık ekran alınmış ve TimeSformer modelinin eğitimi için kullanılmıştır.

Veri setleri ve mimariler aynı olmadığı için ve bu durumda resim ve video arasındaki performans karşılaştırması da adil olmayacağı için, resim veri seti üzerinde TimeSformer mimarisi ve video veri seti üzerinde de ViT mimarisi çalıştırılmıştır. Fakat, ilgili modellere uygun bir şekilde girdi verebilmek için, resim veri seti, tek çerçeveli bir video gibi kullanılmış, video veri seti ise pek çok çerçeveden oluştuğundan dolayı, videonun ortasındaki çerçeve alınarak resim gibi kullanılmıştır. Son olarak, video verisinin felaket sınıflandırmasındaki etkisini görebilmek için, TimeSformer mimarisi, video veri setinden tek çerçeve alınarak mimari bir kez daha çalıştırılmıştır. Tek çerçeve alınırken kullanılan aynı altörnekleme yöntemi uygulanmış ve eğitim için videonun ortasındaki çerçeve kullanılmıştır.

TimeSformer, çok-ekranlı VIDI üzerinde, %76.56 doğruluk elde ederken, tek-ekranlı VIDI üzerinde %67.37 doğruluk elde etmiştir. Bu sonuç, mevcut olduğunda video bilgilerinin kullanılmasının, felaket sınıflandırma performansının iyileştirilmesine yardımcı olduğunu göstermektedir.

Deneylerde, felaket sınıflandırması için son teknoloji mimarilerin, ViT ve TimeSformer, performansı resim ve video veri setlerinde ölçülmüş ve Incidents veri setinin makalesinde kullanılan yaklaşım ile karşılaştırılmıştır. İlgili makalede ResNet-18 mimarisi kullanılmıştır. Bu çalışmada yapılan deneyler sonucunda, ViT ve TimeSformer mimarilerinin, resim tabanlı Incidents veri kümesinde ResNet-18'den daha yüksek doğruluklar elde ettiği gözlemlenmiştir. ResNet-18 tabanlı model %77.3 en iyi bir doğruluk elde ederken, ViT ve TimeSformer sırasıyla %78.5 ve %81.47 en iyi bir doğruluk elde etmiştir.

Ek olarak, ViT ve TimeSformer mimarilerinin performansı, her iki veri kümesinin de tek çerçeveli versiyonları kullanılarak karşılaştırılmıştır. Video tabanlı veri kümesinin tek çerçeveli versiyonunda TimeSformer %67.37 ve ViT %61.78 doğruluk oranı elde etmiştir. Ayrıca, TimeSformer Incidents veri kümesinde %81.47, ViT ise aynı veri setinde %78.5 doğruluk oranı elde etmiştir. TimeSformer'ın hem Incidents hem VIDI veri setinde ViT'den daha yüksek doğruluk elde ettiği gözlemlenmiştir. Ancak, toplanan veri kümesindeki sonuçlar, Incidents veri kümesinde elde edilen sonuçlara kıyasla daha düşük çıkmıştır. Bunun iki ana nedeni öngörülmektedir: (1) resim tabanlı veri kümesi eğitim için daha fazla örnek içerir ve böylece sistemler daha iyi modeller öğrenebilir. (2) Toplanan veri kümesi, sınıflandırma için daha zor örnekler içerir.

Çalışmanın ikinci kısmı negatif örnekler dahil edilerek yapılmıştır. Negatif örnekler ile tüm sınıflar için ikili sınıflandırma modelleri eğitilmiştir. Ana fikir, bir modelin belirli bir videoda o sınıf felaketinin olup olmadığını tespit edebilmesidir. Bu nedenle, 43 felaket çeşidi için 43 farklı model eğitilmiştir. Sonuç olarak, en iyi doğruluk "heyelan" ve "kirli kontamine" sınıfları tarafından elde edilmiştir. "Bloklu" sınıfının tespiti, en düşük doğruluğa sahip olan modeldir.

Son olarak, bu çalışma kapsamında bir sınıflandırma daha yapılmıştır. Diğer sınıflandırma deneylerinden farklı olarak, bu deneyde 44. sınıf olarak negatif örnekler kullanılmıştır. 44. sınıf için hiçbir felaket içermeyen 100 video negatif örneklerin arasından seçilmiştir. Bu deney sonucunda TimeSformer mimarisi %72.18 en iyi bir doğruluk oranını elde etmiştir.

Sonuç olarak, bu çalışmada pek çok sınıf sayısına sahip, çeşitliliği yüksek bir felaket video veri seti sunulmuştur. Sunulan veri setinin pozitif örnekleri kullanılarak sınıflandırma deneyleri yapılmıştır. En yeni mimarilerin video ve resim veri seti üzerindeki performansları karşılaştırılmıştır. Sunulan veri setinin negatif örnekleri de kullanılarak her bir sınıf için ikili sınıflandırma deneyleri yapılmıştır.

1. INTRODUCTION

Nowadays, detection of natural disasters and accidents has become more important. When these kinds of incidents happen, it is important to take action immediately. Humans can easily recognize the type of incidents with a certain confidence. However, using human resources would be difficult in systems which require fast processing. Since more human resources will bring extra cost to the system, using deep learning methods for these challenges will be more efficient.

In this study, it is aimed to classify a disaster by using recent deep learning models. For these experiments, state-of-the-art image and video-based classification techniques are used. In the following sections, the purpose of this study will be explained. Then, the incident classification techniques and the existing datasets in the literature will be shared. Finally, the objectives of this study will be explained.

1.1 Objectives

There are a lot of studies related to incident detection and classification in the literature. Most of them are based on texts and images. Video-based studies are very limited as there are only a few video-based incident datasets. Existing video-based datasets are not diverse and do not include many incident types. To fill this gap, a video-based incident dataset is collected in this study.

There is an existing incident image dataset named Incidents Dataset [1] in literature. It is diverse and it has various class categories. Similar to this dataset, a video-based incidents dataset named VIDI, Video Dataset of Incidents, is presented in this study. The collected video-based dataset complements this image-based dataset since it includes the same 43 classes in the Incidents Dataset. All the videos for the dataset are taken from YouTube[1]. Positive and negative example clips from dataset are shown in Figure 1.1 and Figure 1.2, respectively.

[1] https://www.youtube.com/

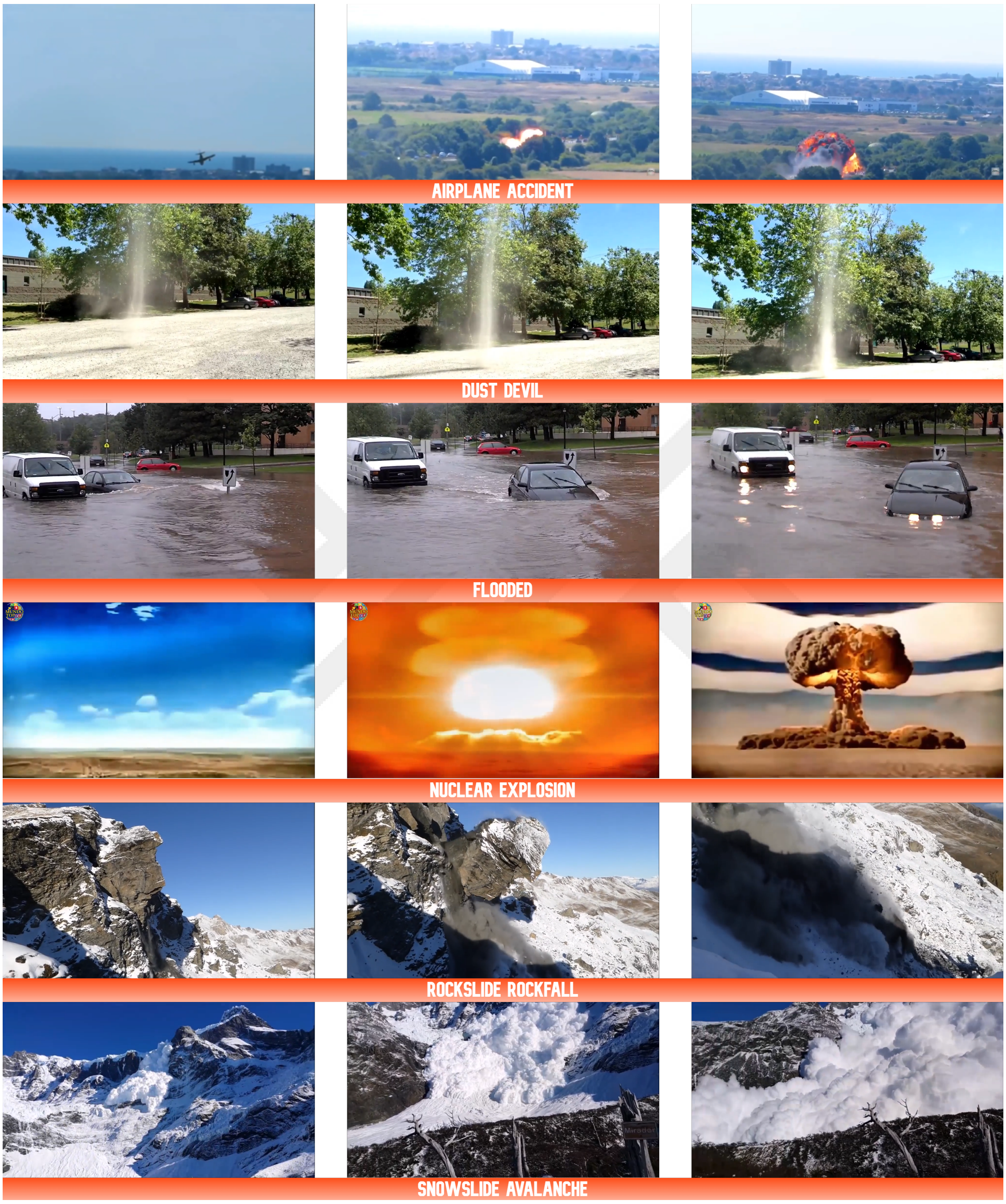

Figure 1.1 : Positive samples from the dataset.

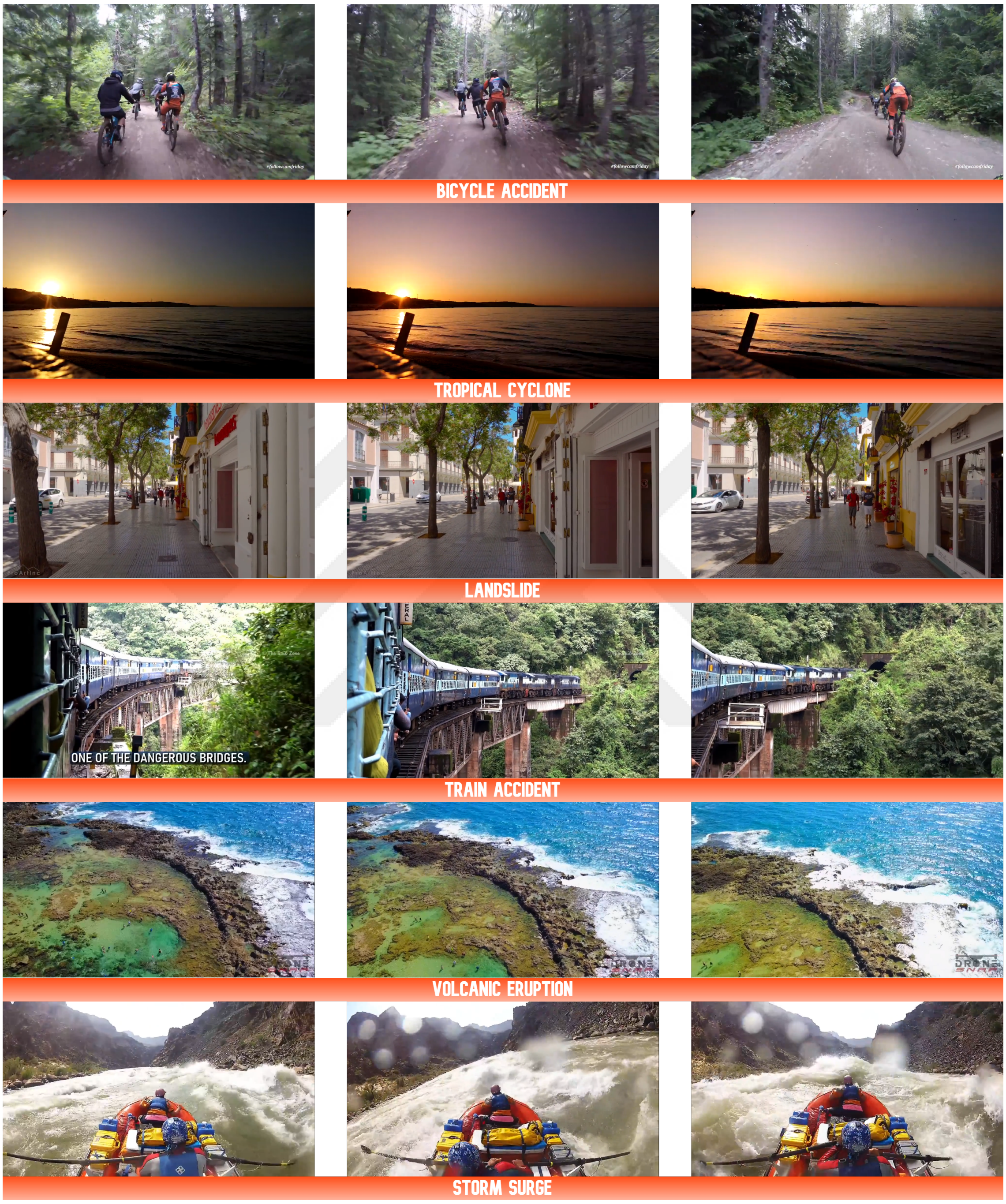

Figure 1.2 : Negative samples from the dataset.

In this study, attempts were made to find answers to several questions. One of the questions was to explore the contribution of video-based information for incident classification. The number of approaches that exploit temporal information is rather limited. One of the main reasons for this is a diverse video dataset with various incident types does not exist. To answer this question, a video-based incidents dataset was collected. After the dataset collection part, single-frame and multiple-frame versions of this video-based dataset are used in the trainings. Based on this scenario, It is aimed to compare the incident classification performance of the recently published video classification architecture on both video (multiple-frame) and image (single-frame) versions of the VIDI.

Another point that is investigated is whether the recently published architectures can give better results than the results that are published in the Incidents Dataset [1]. In their study, ResNet-18 is used to classify incident samples. State-of-the-art video and image classification architectures are run on both the collected dataset, VIDI, and the Incidents Dataset. Vision Transformer and TimeSformer architectures are selected as recently published architectures. By using these architectures, it is demonstrated how well new architectures perform compared to the method that is used in the Incidents Dataset.

Additionally, after collecting all positive and negative incident samples, it is aimed to measure the binary classification performance of each disaster. The goal of this experiment is to determine whether a disaster exists in the video or not.

Finally, in addition to the 43 class classification experiments, 44 class classification experiments are also done for the "no disaster" scenario. For the 44th class, 100 videos that do not include any incidents are selected from the negative samples. The performance of the model is measured when there is a possibility of not having any incidents in the video. To summarize, these are the points that are unraveled in this study with the newly collected dataset and the recent architectures.

1.2 Literature Review

In this part, the studies will be reviewed under three different categories: existing incident datasets, incident classification methods, and transformer-based deep learning methods.

1.2.1 Existing incident datasets

Most of the existing incident datasets are based on texts and images. There are a lot of incident datasets based on social media [4]–[12] and satellite images [13]–[16]. Only one dataset in the literature consists of videos. However, it does not include various types of classes. A list of existing datasets can be seen from Table 1.1. According to this list, it is clear that there is a need for a diverse video-based dataset with the high number of classes. Some of the existing image and video datasets will be examined in the following sections.

Table 1.1 : Statistics of the incident datasets in literature. * is the dataset that is collected in this study.

Dataset Name	# of data	# of class	Type
FIVR	225.960	2	Video
Incidents Dataset	446.684	43	Image
ME17-DIRSM	6.600	2	Image
xBD	22.068	6	Image
VIDI*	8.881	43	Video

1.2.1.1 Image datasets

Most of the image datasets do not have many classes and they do not include various disaster categories. There are also several datasets that are created for a specific disaster. A couple of them are investigated in this study and they inspired this video-based dataset collection study.

The Incidents Dataset is the biggest and the most diverse dataset in the literature. It covers 43 disaster categories and 49 place categories with 446.684 images [1]. Disaster categories are mostly based on natural disasters such as earthquakes, floodings, fires, vehicle accidents, such as car accidents, truck accidents, airplane accidents, and the consequences of these disasters. This dataset consists of positive and negative incident images.

Another natural disaster dataset with images is xBD [17], which contains satellite images of six different types of disasters — fire, hurricane, volcanic eruption, flood, earthquake, and tsunami—. Data is collected from 19 different disaster sites for this dataset. xBD has 22.068 images. Still, it has a limited number of disaster categories and does not cover most of the incidents.

The final image-based disaster dataset is ME17-DIRSM [14]. This dataset contains 6.600 images of flooding incidents. It has positive and negative samples which can be used for binary classification experiments. So, this dataset contains a unique incident type and does not include other types of incidents.

1.2.1.2 Video datasets

In the literature, there is only one video-based dataset [18]. FIVR-200K is a large-scale dataset that mainly searches for fine-grained videos. In this study, Wikipedia[2] is crawled for all events from 2013 to 2017. Then, the events which belong to the "disaster & accident" and "armed conflict & attack" categories are filtered. The filtered events are queried on YouTube. Finally, 225.960 video clips are obtained and labeled with the names of the videos and event headlines.

1.2.2 Incident classification methods

Investigated studies use different methods to classify the incidents. One of them focuses on binary classification experiment [6], and they aim to detect fires from images. To accomplish this task, they also created a dataset. The created dataset contains 114.098 images gathered from Flickr and it is publicly shared. They used four feature detectors — SIFT [19], Color-SIFT [20], SURF [21], Color-SURF [22] —- and they extract the images' keypoints and the descriptors at the key points. By using the k-means algorithm [23], the descriptors are clustered to create the visual vocabulary. Afterwards, a support vector machine (SVM) [24] classifier is trained and the images are classified with this model.

[2] https://en.wikipedia.org/wiki/Main_Page

In another study [7], the main idea was to determine the severity level of the incident damage. There are three classes defined as none, mild, and severe to determine the severity level. They used a dataset that is collected from five different social media sources. The dataset includes 210.000 images. The authors employed three image classification approaches to assess the severity of the incident damage. In their first method, the features are extracted by using Bag-of-Visual-Words [25] and Pyramidal Histogram of Visual Words methods. A linear SVM is trained on these features [7]. Their second method to assess severity is based on a pre-trained VGG-16 model [26]. They used the VGG-16 model as a feature extractor. Their final approach is to fine-tune a pre-trained VGG-16 model to classify the severities of the damage. The best results on all datasets are reached by the fine-tuned deep learning model.

Nogueira et al. aimed to detect floodings from satellite images [14] in another study. They proposed a convolutional neural network based approach. They used a dataset named ME17-DIRSM. This dataset has 6.600 images of flooding incidents and it includes both positive and negative samples in it. Two types of models are used for training. The first model uses dilated convolutions. These convolutions do not change the resolution of the image in the learning process. The second model uses deconvolution networks, the encoder part of the deconvolution networks learns the encoded visual features. The decoder part of the deconvolution networks learns to upsample these features. They trained an SVM classifier on concatenated the outputs of these models.

The final examined study is a method applied to Incidents Dataset. In their study, they did experiments with only positive samples and experiments with both positive and negative samples. A ResNet-18 [27] model is used as a backbone with only positive samples for the classification tasks. The models are trained to jointly recognize the incident types and their place categories. The class-negative loss is used in their training experiments. They achieved remarkable results.

1.2.3 Transformer based deep learning methods

To classify the Incidents Dataset and VIDI as successful as the method that is used in the Incidents Dataset, a literature review has been done on the recent state-of-the-art image and video classification architectures. It is seen that transformer-based architectures are very popular and got successful results [2,3] for different tasks on many datasets [28]–[32]. In this section, two transformer-based architectures will be examined.

1.2.3.1 Vision transformer

Vision transformer [2] is an architecture that is inspired by transformer method used in Natural Language Processing (NLP). It is one of the latest image classification methods which achieves successful results on several datasets. The baseline of the Vision Transformer (ViT) depends on the encoder part of the BERT [33] architecture which does not contain any convolution layers. In Vision Transformer, they showed that a convolutional layer is not necessary for image-based tasks. So, inspired by BERT's technique, they applied transformer blocks to patches of each image sequentially.

Images are split into fixed-size patches and each patch is embedded linearly. The model processes each embedding at the same time in parallel. So, the model needs position information of each patch; therefore, position embeddings are added, and the standard Transformer encoder network is fed with these vectors. Then, the classification is done by the Multi-Layer Perceptron (MLP).

They proposed an architecture by using the transformer encoder part of the network with a different number of transformer blocks. Mainly, three main architectures named Base (12x), Large (24x), and Huge (32x) are proposed. The "Base" and "Large" models are the same as the BERT architecture [33]. The last one consists of 32 transformer encoder blocks. ViT performed better than the ResNet-based models and the best performance is achieved by the "Huge" model [2].

1.2.3.2 TimeSformer

For video processing, there are a lot of studies that use different techniques in the literature. TimeSformer [3] is one of the popular and state-of-the-art video processing architecture. While designing TimeSformer, they are inspired by Vision Transformer [2] and the self-attention mechanism. Similar to ViT, their study also proposed a convolution-free approach. Since it is designed for video processing, the attention mechanism is based on space and time. While ViT architecture, which is designed for images, considers only space, TimeSformer includes both space and time.

In TimeSformer, image patches are extracted from each frame of the video. Each patch is flattened and their positional information and spatial location information are added. Therefore, five new attention mechanisms —space attention, sparse local-global attention, axial attention, joint space-time attention, and divided space-time attention— are created for this video model, and divided space-time attention achieved the best result.

It represents the input video as a time-space sequence. TimeSformer model extracts the features of each patch and the relation between other patches from the input video. Thus, long-term connections are used between distant patches to classify the action better. The results show that the architecture achieved the best video classification accuracy on several action recognition benchmarks [31,32,34].

2. VIDI: A VIDEO DATASET OF INCIDENTS

In this study, a diverse video dataset that includes positive and negative examples of 43 different incident classes is collected. In this part, dataset information and statistics of the dataset, and how the quality checks are done will be explained. Finally, problems that occurred while collecting the dataset will be examined.

2.1 Overview

One of the objectives was to show if temporal information is valuable for incident classification. However, there was no dataset with a large variety of classes in the literature. Therefore, it is aimed to collect a diverse video-based incidents dataset which provides a high number of classes.

The dataset consists of two parts. The first part includes natural disaster videos, such as floodings, earthquakes, fires, vehicle accidents such as truck accidents, bus accidents, and car accidents, and the consequence of these incidents such as burned, collapsed, and damaged. The second part of the dataset contains the videos which do not include the incident of that class. For instance, a negative sample of the "car accident" class should not contain any "car accident" in the video clips.

The categories in this dataset are the same as the ones in the Incidents Dataset. The Incidents Dataset has also been labeled with 49 place categories in addition to incident labels to ensure diversity. However, VIDI does not have any location information labeled. Even though places are not labeled in this dataset, care was taken to collect data from different locations, and this way diversity is ensured.

Another approach to increase diversity is using different languages while searching videos on YouTube. Different languages are used for both positive and negative samples. This approach helped to reach various styles of different cultures. Some of the incidents may differ from location to location or they may occur in a specific region. Used material depends on culture and location, and it may affect the appearance of objects such as cars and buildings. For instance, while wood is used as a construction material in America, limestone is more widely used in the Middle East. In addition, the measures and consequences of different climate events are very different in countries where they are expected and unexpected.

Six widely used languages are preferred while querying videos from YouTube: English, Turkish, Simplified Chinese, Standard Arabic, French, and Spanish. These languages are sometimes not sufficient. In these cases, alternative languages such as German and Hindu have been used.

The language statistics of both negative and positive samples are shared in Table 2.1. It can be seen from Table 2.1, that the languages do not have an equal number of samples. One of the reason why it is not evenly distributed is due to the fact that some disasters are specific to some regions. For instance, derecho mostly exists in North America. So, most of the derecho incident samples are found in English. In order to overcome this problem, the list of countries where this disaster was experienced was searched and queries were made according to the languages used in these regions.

Table 2.1 : Language statistics in the collected dataset.

Language	# of clips in positive samples	# of clips in negative samples
English	2140	1932
French	376	383
Simplified Chinese	343	387
Spanish	560	509
Standard Arabic	280	410
Turkish	452	443
Other	144	282

Neural machine translators are used for the translation of different languages. Most of the time, translations were successful. However, there were some problems with the translation of some incidents. In some cases, the translation of the relevant event did not have the same meaning in the corresponding country. So, this was another issue encountered during data collection on the cross-language translation of disasters. Tropical cyclones can be an example of these situations. When the label “tropical cyclone” is translated into Arabic, the translated word appears to mean the same as a tornado. When possible, we reached out to the native speakers and these problems were resolved by learning its actual meaning in that language.

In addition to the translation problems, it was quite difficult to find videos for some labels. A nuclear explosion can be an example of this problem. Because the videos for the nuclear explosion were recorded in the USA and the USSR only between certain years. After those years, atmospheric tests were banned. So, videos for these events are limited. Therefore, videos that exist on YouTube often contain near duplicates: as the same event was uploaded multiple times from different channels. As a consequence, more than one clip of the same event for this label may be included in this dataset as well.

In the negative sample collection part, videos are found for a label that does not contain the event of that label. For example, videos that are found for negative samples of the "on fire" label, do not include any "on fire" incident. The collected videos are mostly from the videos that the model can be mistaken with that label. Campfires and candles are examples of events that can be confused with the "on fire" label. Another example can be given as flooded. Rivers are used as cases that can be confused with flooding. For vehicle accidents, normal (without an accident) riding videos are put into the dataset.

Additionally, it was hard to find negative videos for some labels that models can confuse. For these cases, positive samples of another label are used as negative samples. To explain with an example, a video that contains an earthquake can be a negative sample for bus accidents. Apart from these events, some videos in negative samples do not include any types of incidents and everything was fine in that video.

Finally, multiple labels are allowed in the dataset. Some videos show more than one type of event; for example, a house that collapsed after an earthquake might be labeled with both earthquake and collapsed, and a car accident would be labeled both traffic accident and damaged. This issue was the last problem experienced while collecting data for the VIDI.

2.2 Labeling and Quality Checks

The first part of the data collection was the labeling and the videos are labeled while searching. All videos are collected manually from YouTube. While querying YouTube, the name of the event categories in different languages is used. In addition, the name of the event categories is combined with different locations. For example, for the "on fire" label, the word "fire" is combined with different locations, and searches were made in YouTube, such as “fire in the building”, and “fire in the forest”. As an example of traffic accidents, traffic accidents that occur in different places such as fields, forests, beaches, and streets were searched.

For eligible clips, the video IDs, the start and end seconds of the corresponding action, the labels, and the language of the video have been extracted and stored. All stages of the actions as pre/action/post-event are included as much as possible. Another point that was taken into consideration was that the videos were only captured from continuous camera view. This means that no scene changes were allowed in the clips. However, both unstable and steady shots were allowed in the clips.

The second part of the data collection was downloading. Videos with low resolutions were downloaded with their own resolutions. On the other hand, video clips with a resolution higher than 720p have been downscaled to 720p and downloaded that way. After that, incidents are cut from downloaded videos by using the start and end seconds of the action.

The last part was the quality check. Downloaded video clips were double-checked. During the quality check, common problems were found. One of the most common issues found was clips being labeled with incorrect incidents. Incidents may not be very clear in the clip and the person who collects the video may misinterpret the clip. So, clips are checked by another person again. Even when the querying is done with the correct keywords, YouTube may return an unrelated event. These problems were solved during this process. One of the other most common problems was labeling the clip with a single label, even if the clip contained multiple events. For example, car crashes under heavy fog should have been labeled with both car accidents and fog, rather than just car accidents. Missing labels are added as new labels.

Since the dataset has a high number of classes, there may be labels that are similar to each other. For this reason, YouTube users may upload similar events, such as the derecho and tornado events, with the wrong titles. In this case, labels are corrected. Similar to this problem, videos that do not contain any incidents related to that label can be uploaded. For example, there may be a video shot in a windy environment and a tropical region, but the clip does not contain any information about the location. In this case, the label of "tropical cyclone" was changed to "derecho". During the data collection phase, there have been multiple iterations of quality checks to detect and correct the issues mentioned above.

2.3 Dataset Statistics

In this part, dataset statistics will be shared. The dataset consists of 8.881 videos in total. 4.534 of the total amount are collected as positive samples. 4.347 of the data are collected as negative samples. There are 43 incident categories and each category has both positive and negative samples. Multiple labels are allowed for positive examples. So, some videos belong to more than one incident class. Therefore, the total number of labels in the dataset is higher than 8.881.

The positive and negative samples of each incident have 100 video clips on average. The number of video clips per class for positive and negative samples are shown in Figure 2.1 and Figure 2.2, respectively.

The video clip duration is limited between 2 to 60 seconds. Video clip duration is ten seconds on average. Short clips have been selected from the broadcasts of news channels as they present a variety of disasters, but they change the camera view frequently. As a result of this, video clips obtained from the news are around two seconds. Therefore, longer clips were collected mostly from documentaries and videos captured with cell phone cameras.

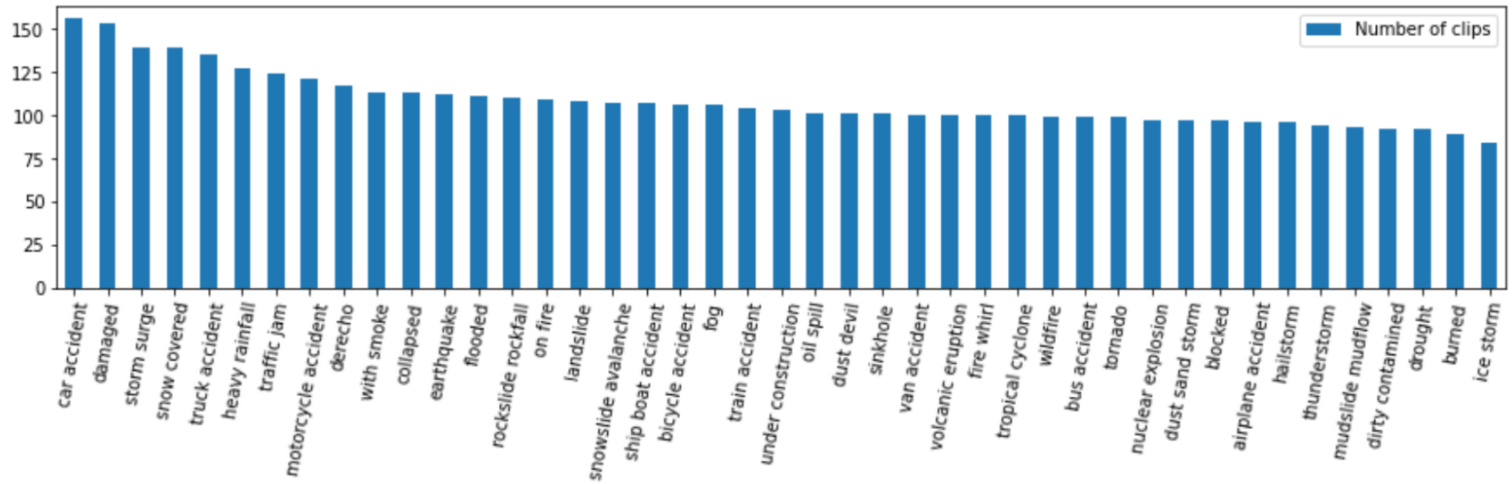

Figure 2.1 : Distribution of positive samples in the dataset.

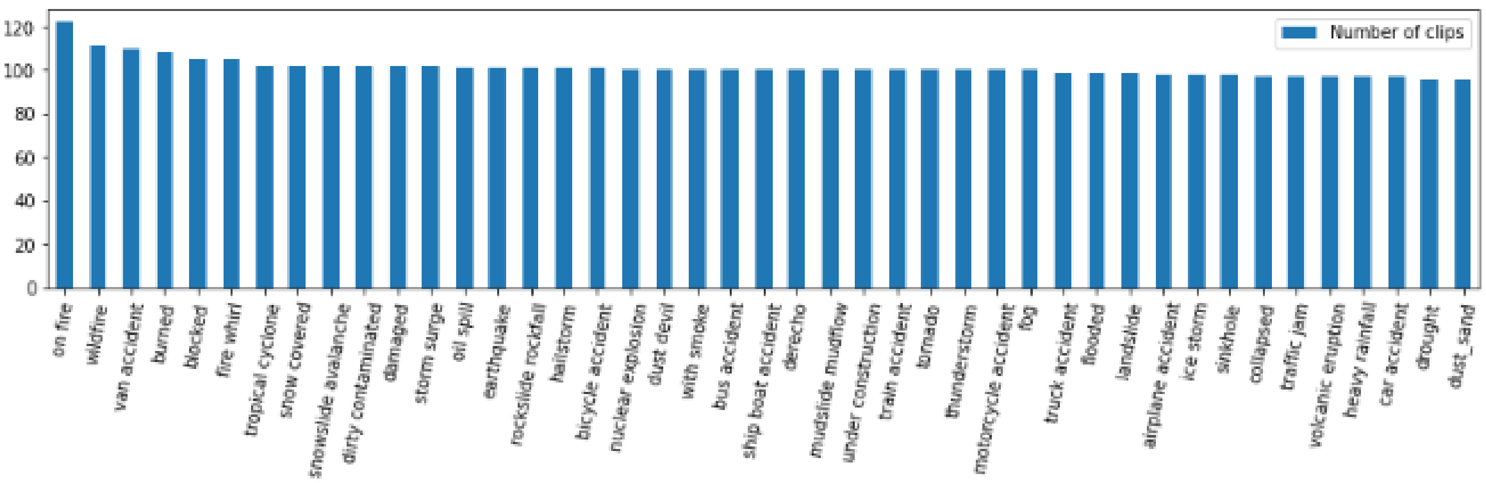

Figure 2.2 : Distribution of negative samples in the dataset.

For the classification experiments, 10% of the class is used for the testing and validation, respectively, and the rest 80% is used for the training. For example, in classification experiments with positive samples in the dataset, 405 and 408 videos were used in the test and validation sets, respectively. If multiple labels are taken into account, it can be said that the test set and the validation set contain 434 and 438 labels, respectively. Videos other than testing and validation were used for training.

For the binary classification experiments, 10% of the classes are used for the testing and validation, respectively, and the rest 80% are used for the training. Since every class has around 100 video clips, approximately ten clips from each class were used for testing and another ten clips are used for validation.

3. METHODS AND IMPLEMENTATION DETAILS

In this part, the architectures used for the experiments will be reviewed and the data processing steps of the architectures will be explained. Then, implementation details for all experiments and hardware information of the machines that the models are run will be shared.

3.1 Architectures

The main purpose of this study is to be able to explore the benefit of processing temporal information for incident classification. Multiple-frame and single-frame versions of the VIDI were needed to compare the performance of video and image classification architectures. After a literature review, two newly designed video and image classification architectures are spotted. These architectures are Vision Transformer (ViT) for image classification and TimeSformer for video classification.

The performance of the most recently published video and image classification architectures is examined for incident classification on video and image datasets. Both architectures used for image and video do not use convolutional layers. Those architectures are inspired by transformer method used in natural language processing and the published results show their superiority for image and video classification tasks.

3.1.1 Vision transformer

The Vision Transformer architecture is designed for image classification. It basically consists of transformer blocks that are used for NLP tasks and does not use any convolutional layers. With the combination of the transformer blocks, ViT architecture is created and it is applied to the images for classification tasks. ViT achieved remarkable accuracy on several benchmark datasets [2].

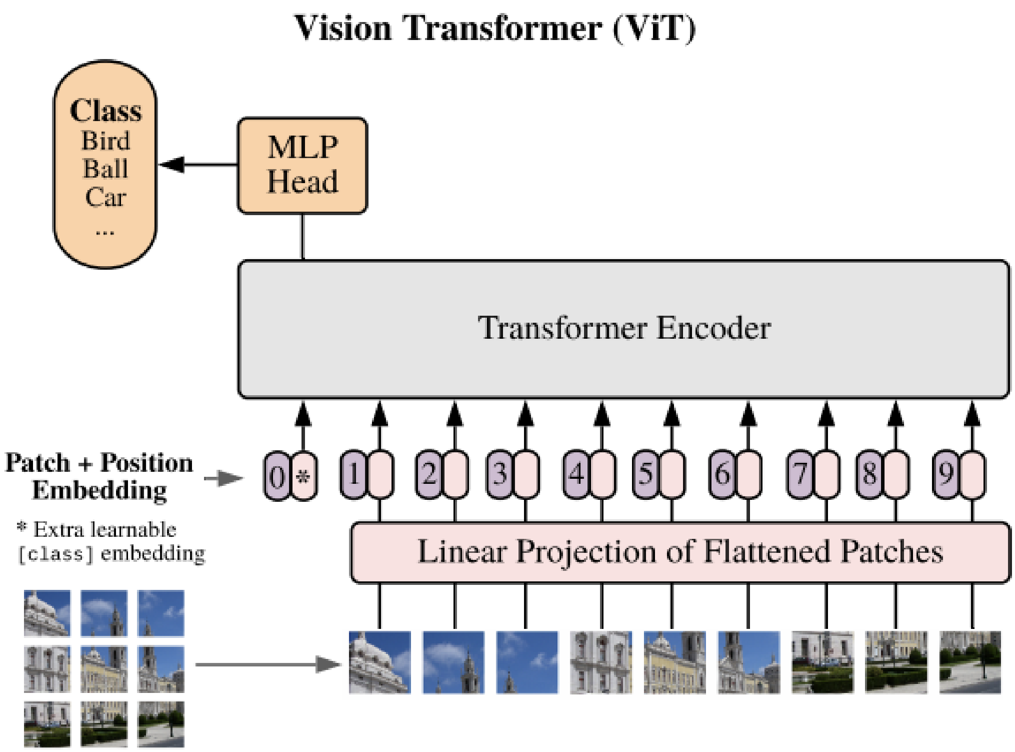

Figure 3.1 : Vision transformer architecture. The figure is taken from [2].

In ViT, first, images are split into patches and each of them is treated as a token as seen in Figure 3.1. The figure is taken from [2]. In the next step, the patches are flattened and reduced to the lower dimension. Since the order of the patches is not known by the model, position information is added to the embedding. Then, standard transformer encoder blocks are fed by the both positional embedding of the patches and flattened patches.

The ViT is first pre-trained on multiple benchmarks such as ImageNet [28], VTAB [29], CIFAR-100 [30], etc. Then, they shared the pre-trained models to be fine-tuned on other specific datasets for classification tasks. Motivated by the accuracy of these pre-trained models on various datasets, this architecture is selected to be used in this study. It is fine-tuned on both Incidents Dataset and the dataset that is collected in this study.

3.1.2 TimeSformer

The TimeSformer model is an architecture that is inspired by Vision Transformer and it is designed for video classification. They adapted new attention mechanisms to the Vision Transformer architecture to process temporal information. Similar to Vision Transformer, TimeSformer also adapts transformer architecture for computer vision tasks and video processing. It consists of a self-attention mechanism and does not contain a convolution layer. It allows learning the spatio-temporal features by using a sequence of patches from a video.

Input clips consist of frames and each frame of the input clips is split into patches. Each patch is embedded linearly. Then, spatial and time positions of the patches are added to the embeddings. After that, different attention mechanisms are applied.

They designed five different attention mechanisms which are shown in Figure 3.2. The figure is taken from [3]. The first one uses only space attention with the residual blocks and it only gets the patches in the same frame. The second one combines both space and time attentions and uses all patches in all frames. The third one applies time attention to the patches in the same frame first. Then, the space attention mechanism is applied to the patches in the same locations but in different frames. The next mechanism focuses on local and global attentions. The last one uses width and height attention with the time attention block. All modules end with the MLP blocks.

In the TimeSformer paper, these five attention mechanisms are run on the K400[1] [32] and SSV2 [35] datasets. Experiments showed that the best result is achieved by the divided Space-Time attention. So, divided Space-Time attention is used for this study.

To sum up, both architectures are pre-trained on the ImageNet and fine-tuned on the Incidents Dataset and the collected video dataset. Both architectures are designed convolution-free and are based on the transformer encoder blocks.

[1] https://bit.ly/3I9TgZi

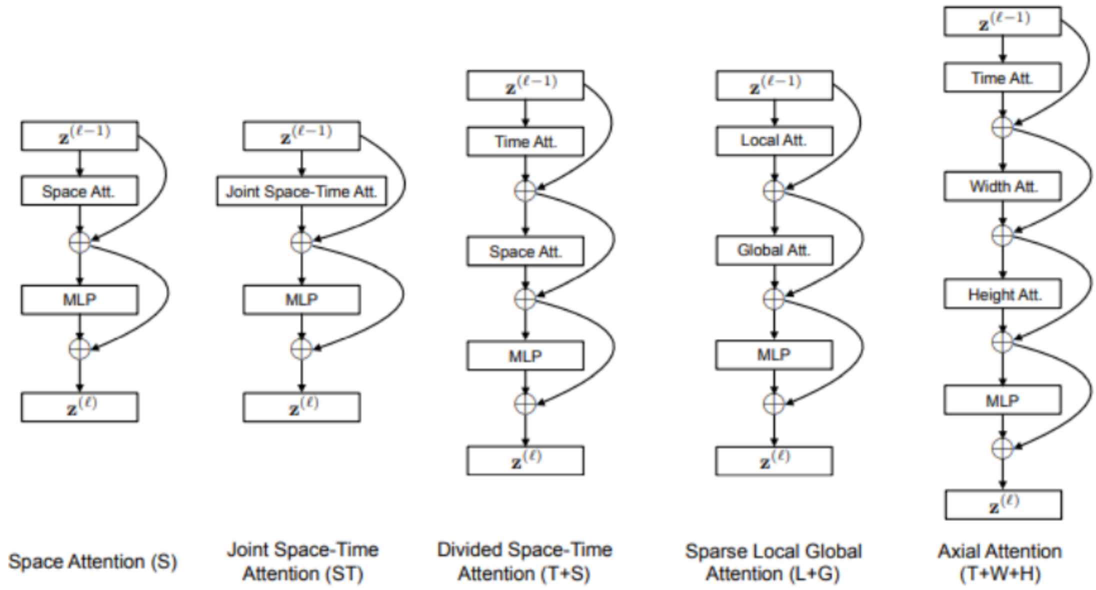

Figure 3.2 : TimeSformer architecture. The figure is taken from [3].

3.2 Implementation Details

ViT and TimeSformer architectures are trained and several experiments are conducted on VIDI and Incidents Dataset. All the parameters used in the experiments are given in Table 3.1. In this study, the PyTorch[2] framework is used for the implementation.

First, the experimental setup is a closed set incident classification task where all positive samples are used. ViT is fine-tuned on Incidents Dataset to generate the incident image-based classification model. Then, TimeSformer is fine-tuned on the collected dataset to obtain the incident video-based classification model. Eight frames are sampled from each video in the collected dataset. Since experiments are run on different datasets with different architectures, it would not be fair to compare image classification performance to video classification performance. To have a fair comparison and see the effect of temporal information, video classification architecture (TimeSformer) is trained on Incidents Dataset which does not include any temporal information. However, input data was not suitable for these models. To overcome this issue, the image dataset is treated as a video with a single-frame. Additionally, image

[2] https://pytorch.org/

Table 3.1 : Implementation parameters.

Experiments	Optimizer	Loss Fn	Base Learning Rate	LR Shrinkage	# of Frames
ViT on Incidents Dataset	SGD	CE Loss	0.001	lambdaLR	1
TimeSformer on Incidents Dataset	Adam	CE Loss	0.005	-	1
TimeSformer on VIDI	SGD	BCE with Logits	0.01	MultiStepLR	8
TimeSformer on VIDI*	SGD	BCE with Logits	0.01	MultiStepLR	1
ViT on VIDI	SGD	BCE with Logits	0.25	ReduceLROnPlateau	1
Binary classification on VIDI*	SGD	CE Loss	0.01	MultiStepLR	8
44 classes on VIDI*	SGD	BCE with Logits	0.01	-	8

classification architecture (Vision Transformer) is trained on VIDI. As the video dataset consists of multiple frames, the middle frame of the input video is taken. The selected frame is treated as an image. To be able to compare the results on the TimeSformer, the architecture is performed one more time with the single-frame version of VIDI. The same sub-sampling method is applied and the middle frame is used for the training.

Another experiment is performed only on the video dataset. In this experiment, it is aimed to detect whether a specific incident exists in the video or not. For this purpose, 43 models are trained for each incident class and their parameters are also shared in Table 3.1 with the name "Binary classification on VIDI".

Finally, one more classification experiment is done on the video dataset. In addition to the 43 incident classes, one more class is added. This class is added for the videos which do not contain any incidents. To create this class, negative samples are used. From all negative samples, 100 videos that do not include any disaster are sampled. That is, the 44th class is used for this case. Implementation parameters are given in the Table 3.1 under the name of "44 classes on VIDI".

B_16_imagenet1k[3] is used for Vision Transformer as the pre-trained model. The divided space-time model is used for TimeSformer. B_16_imagenet1k was trained on ImageNet-21k [28] image dataset and the divided space-time model was trained on the K400 [32] and K600[4] [31] video datasets.

Hyperparameter optimizations are done on the validation set. According to the results, first experimental setup, ViT on Incidents Dataset, used stochastic gradient descent optimization method [36] which lacks adaptive learning rate and only iterates through momentum. The cross-entropy loss [37] was preferred in this experiment. 0.001 learning rate is set as a base learning rate and it is updated during training by Lambda LR method. The learning rate is updated according to the current epoch number in the Lambda LR method. The learning rate is decreased in every three epochs in these experiments.

[3]https://bit.ly/3CMzSkb
[4]https://bit.ly/36iq1Xi

The second experimental setup uses the Adam optimizer [38] which is using momentum along with learning rate, and iterating over and over on it to converge to the best possible solution. It is used for the fine-tuning process of the TimeSformer model on the Incidents Dataset. The cross-entropy loss [37] is selected and its base learning rate is set to 0.005.

The third experimental setup, TimeSformer on VIDI, used stochastic gradient descent optimization method. Since multi-labeling is allowed and a video might belong to more than one incident on the collected dataset. Binary-cross entropy with logits loss is chosen for the classification experiments. Base learning rate is set to 0.01. Multi-step LR method is used and the learning rate is reduced in every five epochs. TimeSformer on VIDI* experimental setup is also using the same parameters with the TimeSformer on VIDI experimental setup. Only difference is number of parameters. While TimeSformer on VIDI uses one frame, TimeSformer on VIDI* uses eight frames.

The next experimental setup is ViT on VIDI. In this setup, stochastic gradient descent optimization method and binary-cross entropy with logits loss is used and base learning rate is set to 0.25. ReduceLROnPlateau method which decreasing the learning rate if the changes in the loss are less than a given threshold during two epochs is used as a learning rate scheduler.

In the binary classification experiments, stochastic gradient descent optimization method and cross-entropy loss is used and base learning rate is set to 0.01. To adapt the learning rate, multi-step LR method is used.

The last experimental setup is designed for open-set classification. Similar to other classification experiments on VIDI, stochastic gradient descent optimization method and binary-cross entropy loss is used. 0.001 learning rate is set as a base learning rate.

3.3 Hardware Information

In this part, the hardware and software features of the computers on which experiments are performed will be given. two different computers and google compute engines are used for training.

- Computer 1
  - CPU : Intel i7 8700K @ 3.70Ghz (6 cores, 12 processors)
  - RAM : 32768 MB
  - GPU : NVIDIA GeForce GTX 1070 Ti
  - VRAM : 8079 MB
  - OS : Windows 10 Education 64-bit (v. 21H2)
  - Python 3.8.7 (64-bit) with the following libraries: PyTorch (1.7.1+cu110), NumPy,
  - pandas, pickle
  - NVIDIA (R) Cuda compiler driver (V10.0.130)
  - C# 8.0 with .Net Core 3.1 Framework
  - Visual Studio Code (1.63.2)
- Computer 2
  - CPU : Intel i7 5820K @ 3.30Ghz (6 cores, 12 processors)
  - RAM : 32068 MB
  - GPU1 : NVIDIA GeForce GTX TITAN X
  - GPU2 : NVIDIA GeForce GTX TITAN X
  - VRAM1 : 12209 MB
  - VRAM2 : 12212 MB
  - OS : Ubuntu 16.04.7 LTS (4.15.0-142-generic)
  - Python 3.9.1 (64-bit) with the following libraries: PyTorch (1.9.1+cu102), NumPy,
  - pandas, pickle
  - NVIDIA (R) Cuda compiler driver (V11.2.67)
  - Jupyter client (7.0.6) on Jupyter core (4.9.1) with IPython (7.29.0) and ipykernel (5.3.4)

- Google Compute Engine:
  - CPU : Intel(R) Xeon(R) CPU @ 2.20GHz
  - RAM : 13298 MB
  - OS : Ubuntu 18.04.5 LTS (Bionic Beaver)
  - Python 3.7.13 (64-bit) with the following libraries: NumPy,
  - pandas

Computer 1 is used in the classification experiments with the positive samples. Computer 2 is used in the detection experiments with the negative samples and 44 class classification experiments. Finally, Google Colab[5] is used as a compute engine and it is mainly used to develop training and utility codes before actual training started on other computers.

[5] https://colab.research.google.com/

4. EXPERIMENTS

In this part, the experiments will be explained in detail and the performance of the models will be compared. After all experiments are completed, error analyses are done on the misclassified predictions.

Answers to the following questions were sought in the experiments. The first one is "How the temporal information affects performance?". "Is the performance of the recent architectures better on the Incidents Dataset?" is the second question. Then, "How well video classification architectures perform in binary classification experiments for each disaster?" is the next question. The last question is, "What would be the effect of formalising the problem as an open set classification by adding one more class for the ones which do not contain any incidents on the video classification experiments?".

Experiments are done on both Incidents Dataset and the collected dataset. Incidents Dataset is already split into three parts: training, validation, and test. Only class-positive samples are used from the Incidents Dataset. While there are 312.742 images in the training set of the Incidents Dataset, there are 17.357 images in the validation set and 17.255 images in the test set. In the classification experiments on VIDI, only the positive samples are used. 3.721 videos of the dataset are allocated for training, 408 videos are allocated for validation, and 405 videos are allocated for testing. Please note that both datasets have the same incident categories and can include multiple labels per image/video, that is, an image or video might belong to more than one incident category.

For the binary classification experiments, both positive and negative samples are used. 10% of the dataset is allocated for the testing and validation part, respectively, and the rest is used for the training.

4.1 Evaluation Metrics

Top-1 and top-5 accuracies are used for the evaluation. Top-1 accuracy is a performance measurement concept that expects the model to give the correct answer within the first most likely prediction. Top-5 accuracy is also a similar performance measurement concept that expects the model to give the correct answer within the five most likely predictions.

4.2 Results

Two architectures are trained using only positive samples of both datasets to perform the incident classification in the first part of the experiments. Different architectures are performed on the video and image formats of both datasets.

Results of the experiments use only positive samples are shown in Table 4.1. On each run, the test videos are predicted three times in the testing phase with different crops. The summation method is used to combine the probabilities of each prediction of a clip.

Table 4.1 : Top-1 and Top-5 accuracies (%) of different architectures on the different datasets. *Using multiple frames. Single frame is used in the rest of the experiments. †Result published in [1].

Architecture	Dataset	Top-1 Acc	Top-5 Acc
ResNet-18†	Incidents Dataset	77.3	95.9
ViT	Incidents Dataset	78.5	96.33
TimeSformer	Incidents Dataset	**81.47**	**96.95**
ViT	VIDI	61.78	86.78
TimeSformer	VIDI	67.37	90.59
TimeSformer	VIDI*	**76.56**	**96.51**

One of the objectives is to explore whether temporal information is valuable for incident classification. To be able to address this question, TimeSformer architecture is fine-tuned on both multiple-frame and single-frame versions of VIDI. While the single-frame version achieves 67.37% accuracy, Fine-tuned TimeSformer architecture obtained 76.56% accuracy on the multiple-frames version. This comparison shows that temporal information helps to get better classification performance.

The performance of the recently published ViT and TimeSformer architectures for incident classification experiments is measured. Then, the results are compared with the results in [1], which uses ResNet-18 architecture. As seen in Table 4.1, recent and convolution-free architectures performed better than the ResNet-18 architecture. According to the published results in [1], the accuracy of classification with positive samples is 77.3%. In this experiment, 78.5% and 81.47% top-1 accuracies are obtained by ViT and TimeSformer architectures, respectively, on the Incidents Dataset.

Furthermore, a performance comparison between ViT and TimeSformer architecture is done on the single-frame version of VIDI and Incidents Dataset with these experiments. It is seen that TimeSformer outperformed ViT on both datasets. However, when the results on the different datasets are compared, the accuracy on the collected dataset is lower than the accuracy on the Incidents Dataset. There could be two reasons for this. The first one is the amount of data. There are more samples in the Incidents Dataset. So, it might allow to learn better models. The second reason might be the difficulty of the samples. The video dataset might contain more complex samples than Incidents Dataset.

The rest of the experiments are done with both positive and negative samples. The first experiment with them is a 44 class classification. One more class is added to the 43 class incident classification experiment. The added class stands for the case which does not include any incident. While the TimeSformer architecture achieves 76.56% top-1 and 96.51% top-5 accuracy on 43 class classification experiments, it achieves 72.18% top-1 and 93.29% top-5 accuracy on 44 class multi-frame classification experiments according to results in Table 4.2. The results of 44 class classification are lower than the results of 43 class classification experiments. It can be said that it has achieved acceptable results considering that the number of classes has been increased.

Table 4.2 : Top-1 and Top-5 accuracies (%) of 44 class classification experiment in the collected dataset. *Using multiple frames.

Architecture	Dataset	Top-1 Acc	Top-5 Acc
TimeSformer	VIDI*	72.18	93.29

Finally, binary classification experiments are done for each class with the TimeSformer architecture on the multi-frame version of the VIDI. 43 models are trained for 43 classes. The model of each class is responsible for whether that disaster is in the video. Some of the classes got higher scores, while some of them got lower scores. Accuracy per class can be seen in Figure 4.1. According to this figure, "landslide", "dirty contamined", "derecho", "tornado", "fog", "dust devil", "drought", "volcanic eruption", "snow-covered", and "on fire" are the ones that reach over 95% accuracy. However, the following classes are the ones with lower than 85% accuracy: "dust sand storm", "oil spill", "sinkhole", "bicycle accident", "fire whirl", "mudslide mudflow", "under construction", "collapsed", "burned", and "ship boat accident". Lastly, "blocked" has the lowest accuracy with 65.62%.

4.3 Error Analysis

After experiments are done, the wrong predictions in the classification experiments are analyzed. A couple of examples of wrong predictions from 43 class classification experiments can be seen in Figure 4.2. According to these observations, it is seen that the prediction of the video clips is very similar to the visual content in the video and it can be easily predicted wrong. The false prediction matrix of 43 class classification experiments is given in Figure 4.3. From this figure, it can be seen which label is mixed with which label by the model in the multi-frame experiment and how many times it made a mistake. "Ice storm & snow-covered", "landslide & rockslide rockfall", and "flooded & storm surge" are classes that are misclassified mostly.

Furthermore, accuracy per class in the 43 class classification experiments is calculated and the results are shown in Figure 4.4. "Dirty contamined", "bicycle accident", "nuclear explosion", "airplane accident", "traffic jam", "wildfire", "tornado", "dust devil", "dust sand storm", and "snow-covered" are the classes that are predicted with higher accuracy than the other classes on 43 class classification experiments. In the same experiment, it is seen that the lowest prediction accuracy belongs to the "landslide" class. According to Figure 4.3, most of the mistakes are between the different vehicle accidents. The least successfully classified label can also be seen in Figure 4.4.

Error analyses are also done for the 44 class classification experiments which include negative samples. As shown in Figure 4.5, although the true class of the video is negative and does not include any disaster, it is predicted as there is an incident in the video. For example, a video is predicted as "on fire", even though campfire is not an incident. In contrast, some of the videos that include incidents are predicted as there are no incidents in this video. The samples that are mixed with the "negative" class are shown in Figure 4.6. According to examples shown in Figure 4.6, several videos are predicted as "on fire" and "van accident", even though they do not include any incident. Furthermore, "airplane accident" and "wildfire" samples are predicted as "negative".

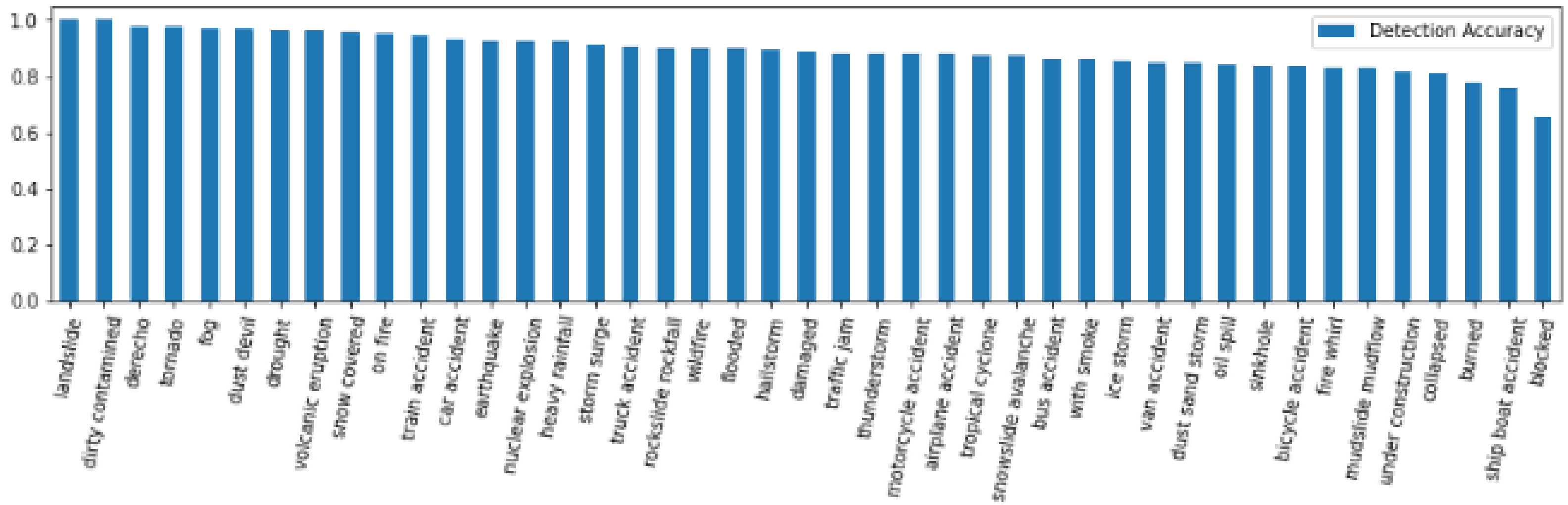

Figure 4.1 : Accuracy of the models for each class in the binary classification experiments.

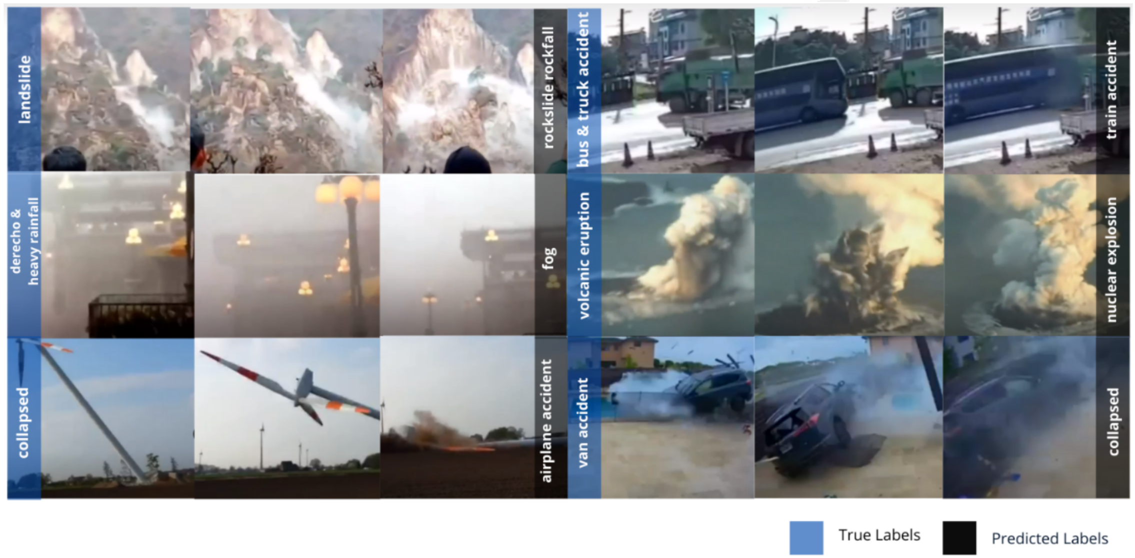

Figure 4.2 : False classification examples from the 43 class classification predictions. Three frames of each video are demonstrated with the correct labels on its left, and the predicted labels on its right.

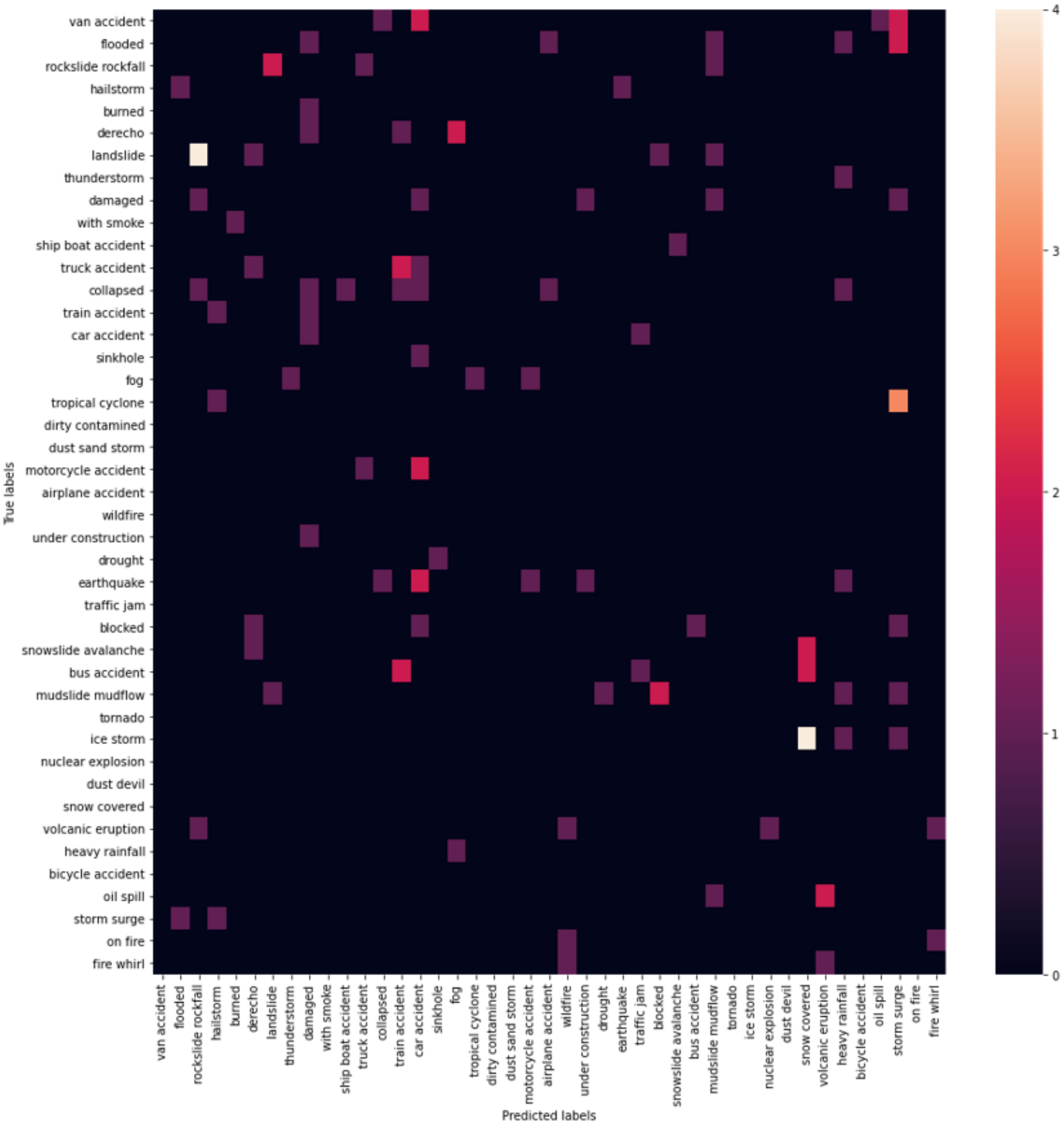

Figure 4.3 : A matrix showing which video was predicted by which false label in 43 class classification experiment.

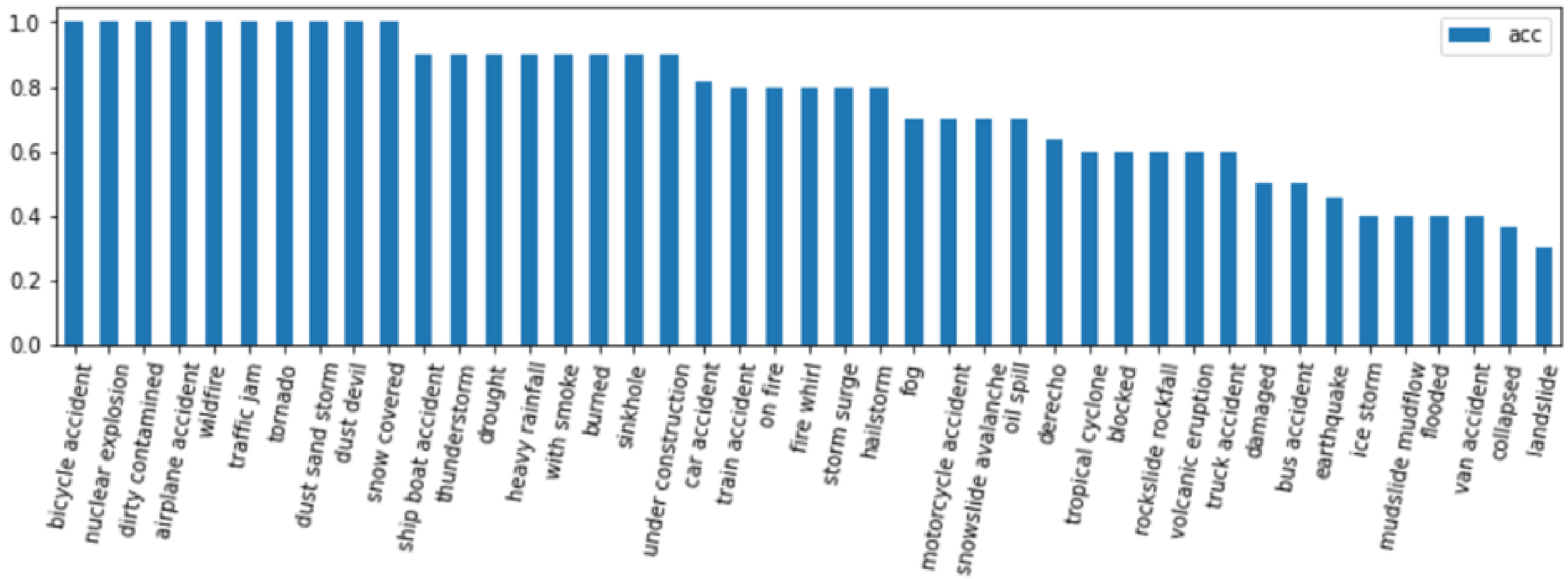

Figure 4.4 : The performance of each class in 43 class classification experiment.

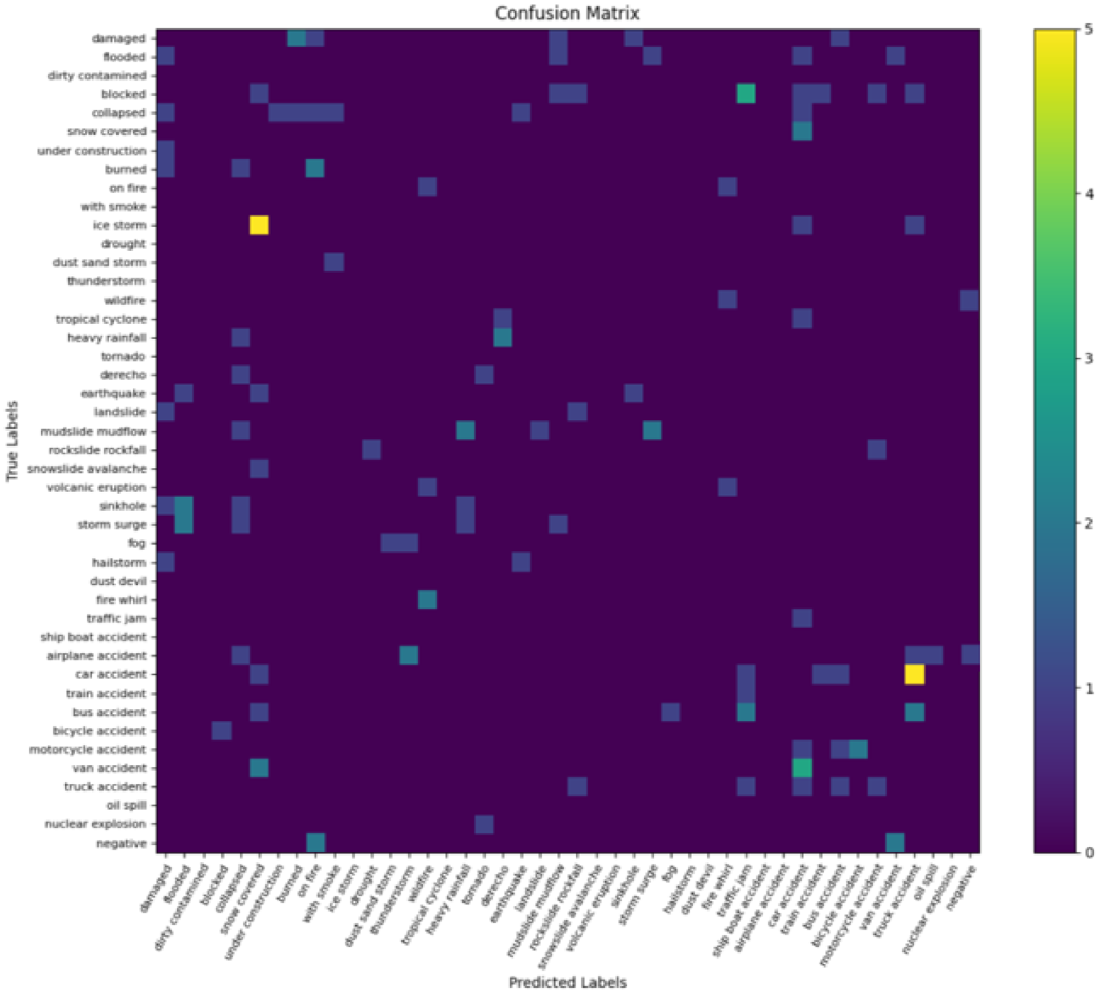

Figure 4.5 : A matrix showing which video was predicted by which false label in 44 class classification experiment.

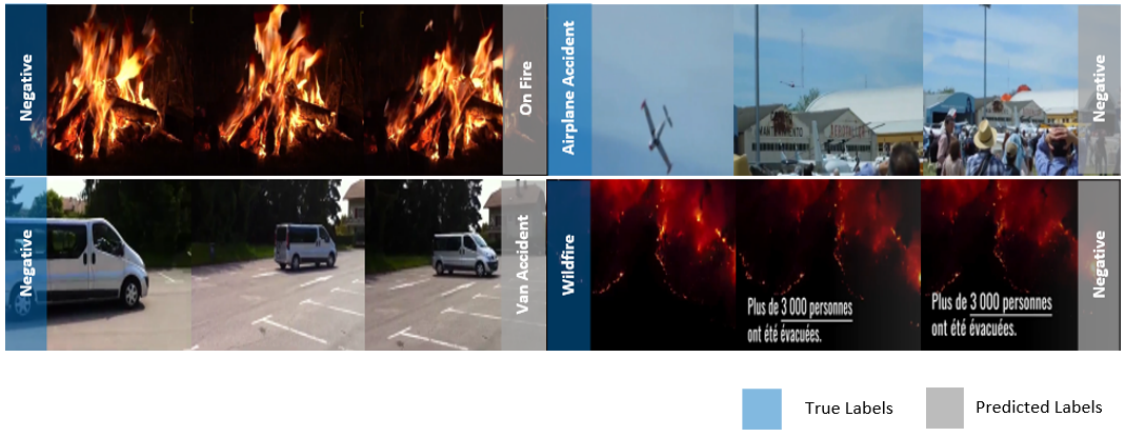

Figure 4.6 : False classification examples from the 44 class classification predictions. Three frames of each video are demonstrated with the correct labels on its left, and the predicted labels on its right.

5. CONCLUSION AND FUTURE WORK

The presented work aimed to increase disaster classification performance in smart systems. Motivated by this, whether the temporal information has an important role in incident classification was investigated. Most of the questions related to incident classification are resolved with this study.

A dataset that can provide both temporal and spatial information was needed to be able to have a fair comparison. There is an existing diverse and multi-classes incident image-based dataset named Incidents Dataset. However, there was no diverse incident video-based dataset with various classes. So, to fulfill this need, a video-based dataset is collected. The collected dataset also has the same number of classes and it complements the Incidents Dataset.

In this study, several experiments are done with the collected dataset. The state-of-the-art image and video classification architectures are fine-tuned on Multi-frame and single-frame versions of the collected dataset. Additionally, these recent architectures are run on the previously published dataset. Finally, binary classification experiments are done for each class. According to the experiments, the results are promising and provide baselines for future studies.

To sum up, the community was a lacking publicly available diverse video-based incidents dataset with the high number of classes. The presented dataset is very useful for the related research community. Also, the initial incident classification experiments of this study present promising results. Furthermore, it proves that temporal information is valuable for incident classification. Additionally, recent architectures achieved better results than previously published results. Finally, binary classification experiments show how well the model performed for each class.

So far, multi-frame experiments are performed with a certain number of frames. The influence of the number of frames can be investigated in the next experiments. In this study, eight frames are extracted from the video, additional experiments can be conducted to see whether increasing the number of frames will affect the performance.

Furthermore, the model trained on the VIDI can be tested on the test set of Incidents Dataset and its classification performance can be measured. Similar to this experiment, the model trained on the Incidents Dataset can be tested for classification experiments on the test set of VIDI.

For the 44 class classification experiments, 100 videos that do not include any disaster are selected from negative samples. In the future experiments, all videos that do not include any disaster from negative samples can be used for the 44 class classification experiments. Additionally, binary classification experiments can be done for "incident" and "no incident" scenarios with all samples.

TimeSformer achieved higher accuracy than ViT on single-frame experiments. However, the single-frame version does not contain temporal information, it is expected to see similar accuracy on both architectures. Therefore, the difference between TimeSformer and ViT architectures on single-frame experiments can be investigated.

Confusion matrices of 43 class classification experiments between published results of Incidents Dataset and the results of this study can be compared to see which classes are predicted better in which experiment. Finally, the amount of positive and negative samples in the dataset can be increased in the future.

REFERENCES

[1] **Weber, E.**, **Marzo, N.**, **Papadopoulos, D.**, **Biswas, A.**, **Lapedriza,** , **Ofli, F.**, **Imran, M. and Torralba, A.** (2020). Detecting Natural Disasters, Damage, and Incidents in the Wild, ECCV, pp.331–350.

[2] **Kolesnikov, A.**, **Dosovitskiy, A.**, **Weissenborn, D.**, **Heigold, G.**, **Uszkoreit, J.**, **Beyer, L.**, **Minderer, M.**, **Dehghani, M.**, **Houlsby, N.**, **Gelly, S.** *et al.* (2021). An image is worth 16x16 words: Transformers for image recognition at scale, *ICLR*.

[3] **Bertasius, G.**, **Wang, H. and Torresani, L.** (2021). Is Space-Time Attention All You Need for Video Understanding?, *Proceedings of the International Conference on Machine Learning (ICML)*.

[4] **Peters, R. and de Albuquerque, J.P.** (2015). Investigating images as indicators for relevant social media messages in disaster management., *ISCRAM*.

[5] **Chen, T.**, **Lu, D.**, **Kan, M.Y. and Cui, P.** (2013). Understanding and classifying image tweets, *Proceedings of the 21st ACM International Conference on Multimedia*, pp.781–784.

[6] **Daly, S. and Thom, J.A.** (2016). Mining and Classifying Image Posts on Social Media to Analyse Fires., *ISCRAM*, pp.1–14.

[7] **Nguyen, D.T.**, **Ofli, F.**, **Imran, M. and Mitra, P.** (2017). Damage assessment from social media imagery data during disasters, *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp.569–576.

[8] **Nia, K.R. and Mori, G.** (2017). Building damage assessment using deep learning and ground-level image data, *2017 14th Conference on Computer and Robot Vision (CRV)*, IEEE, pp.95–102.

[9] **Alam, F.**, **Ofli, F. and Imran, M.** (2018). Processing social media images by combining human and machine computing during crises, *International Journal of Human–Computer Interaction*, *34*(4), 311–327.

[10] **Li, X.**, **Caragea, D.**, **Caragea, C.**, **Imran, M. and Ofli, F.** (2019). Identifying disaster damage images using a domain adaptation approach, *Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management*.

[11] **Pouyanfar, S.**, **Tao, Y.**, **Sadiq, S.**, **Tian, H.**, **Tu, Y.**, **Wang, T.**, **Chen, S.C. and Shyu, M.L.** (2019). Unconstrained flood event detection using adversarial data augmentation, *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp.155–159.

[12] **Abavisani, M.**, **Wu, L.**, **Hu, S.**, **Tetreault, J. and Jaimes, A.** (2020). Multimodal categorization of crisis events in social media, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.14679–14689.

[13] **Gueguen, L. and Hamid, R.** (2015). Large-scale damage detection using satellite imagery, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1321–1328.

[14] **Nogueira, K.**, **Fadel, S.G.**, **Dourado, C.**, **de O. Werneck, R.**, **Muñoz, J.A.V.**, **Penatti, O.A.B.**, **Calumby, R.T.**, **Li, L.T.**, **dos Santos, J.A. and Torres, R.d.S.** (2018). Exploiting ConvNet Diversity for Flooding Identification, *IEEE Geoscience and Remote Sensing Letters*, *15*(9), 1446–1450.

[15] **Li, Y.**, **Ye, S. and Bartoli, I.** (2018). Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning, *Journal of Applied Remote Sensing*, *12*(4), 045008.

[16] **Radke, D.**, **Hessler, A. and Ellsworth, D.** (2019). FireCast: Leveraging Deep Learning to Predict Wildfire Spread., *IJCAI*, pp.4575–4581.

[17] **Gupta, R.**, **Goodman, B.**, **Patel, N.**, **Hosfelt, R.**, **Sajeev, S.**, **Heim, E.**, **Doshi, J.**, **Lucas, K.**, **Choset, H. and Gaston, M.** (2019). Creating xBD: A dataset for assessing building damage from satellite imagery, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp.10–17.

[18] **Kordopatis-Zilos, G.**, **Papadopoulos, S.**, **Patras, I. and Kompatsiaris, I.** (2019). FIVR: Fine-Grained Incident Video Retrieval, *IEEE Transactions on Multimedia*, *21*(10), 2638–2652, `https://doi.org/10.1109/TMM.2019.2905741`.

[19] **Lowe, D.** (1999). Object recognition from local scale-invariant features, *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pp.1150–1157 vol.2.

[20] **Kumari, M.S. and Shekar, B.** (2010). Color-SIFT model: a robust and an accurate shot boundary detection algorithm, *Second International Conference on Digital Image Processing*, volume7546, International Society for Optics and Photonics, p.75463S.

[21] **Bay, H.**, **Tuytelaars, T. and Gool, L.V.** (2006). SURF: Speeded up robust features, *European Conference on Computer Vision*, Springer, pp.404–417.

[22] **Fan, P.**, **Men, A.**, **Chen, M. and Yang, B.** (2009). Color-SURF: A surf descriptor with local kernel color histograms, *2009 IEEE International Conference on Network Infrastructure and Digital Content*, pp.726–730.

[23] **Hart, P.E.**, **Stork, D.G. and Duda, R.O.** (2000). *Pattern classification*, Wiley Hoboken.

[24] **Cortes, C. and Vapnik, V.** (1995). Support-vector networks, *Machine Learning*, *20*(3), 273–297.

[25] **Csurka, G.**, **Dance, C.**, **Fan, L.**, **Willamowski, J. and Bray, C.** (2004). Visual categorization with bags of keypoints, *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, Prague, pp.1–2.

[26] **Simonyan, K. and Zisserman, A.** (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition, *ICLR*, *abs/1409.1556*.

[27] **He, K.**, **Zhang, X.**, **Ren, S. and Sun, J.** (2016). Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

[28] **Deng, J.**, **Dong, W.**, **Socher, R.**, **Li, L.J.**, **Li, K. and Li, F.F.** (2009). ImageNet: a Large-Scale Hierarchical Image Database, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255.

[29] **Zhai, X.**, **Puigcerver, J.**, **Kolesnikov, A.**, **Ruyssen, P.**, **Riquelme, C.**, **Lucic, M.**, **Djolonga, J.**, **Pinto, A.S.**, **Neumann, M.**, **Dosovitskiy, A.** *et al.* (2019). A large-scale study of representation learning with the visual task adaptation benchmark, *arXiv preprint arXiv:1910.04867*.

[30] **Krizhevsky, A.**, **Hinton, G.** *et al.* (2009). Learning multiple layers of features from tiny images.

[31] **Carreira, J.**, **Noland, E.**, **Banki-Horvath, A.**, **Hillier, C. and Zisserman, A.** (2018). A Short Note about Kinetics-600, *arXiv e-prints*, arXiv–1808.

[32] **Carreira, J. and Zisserman, A.** (2017). Quo vadis, action recognition? a new model and the kinetics dataset, *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.6299–6308.

[33] **Devlin, J.**, **Chang, M.W.**, **Lee, K. and Toutanova, K.** (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp.4171–4186, `https://aclanthology.org/N19-1423`.

[34] **Goyal, R.**, **Ebrahimi Kahou, S.**, **Michalski, V.**, **Materzynska, J.**, **Westphal, S.**, **Kim, H.**, **Haenel, V.**, **Fruend, I.**, **Yianilos, P.**, **Mueller-Freitag, M.** *et al.* (2017). The "something something" video database for learning and

evaluating visual common sense, *Proceedings of the IEEE International Conference on Computer Vision*, pp.5842–5850.

[35] **Goyal, R.**, **Ebrahimi Kahou, S.**, **Michalski, V.**, **Materzynska, J.**, **Westphal, S.**, **Kim, H.**, **Haenel, V.**, **Fruend, I.**, **Yianilos, P.**, **Mueller-Freitag, M.** *et al.* (2017). The" something something" video database for learning and evaluating visual common sense, *Proceedings of the IEEE international conference on computer vision*, pp.5842–5850.

[36] **Sutskever, I.**, **Martens, J.**, **Dahl, G. and Hinton, G.** (2013). On the importance of initialization and momentum in deep learning, *International conference on machine learning*, PMLR, pp.1139–1147.

[37] **Bishop, C.M. and Nasrabadi, N.M.** (2006). *Pattern recognition and machine learning*, volume 4, Springer.

[38] **Kingma, D.P. and Ba, J.** (2015). Adam: A Method for Stochastic Optimization. ICLR. 2015, *9*.

CURRICULUM VITAE

Name Surname: Duygu SESVER

EDUCATION:

- **B.Sc.:** 2018, Yıldız Technical University, Faculty of Electrical & Electronics Engineering, Computer Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- April 2020 - Present, Machine Learning Engineer at Lely Turkey
- August 2018 - March 2020, Software Developer at AVL Turkey
- March 2018 - July 2018, Project Intern at Cybersoft

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Sesver D.**, Gençoğlu A. E., Yıldız Ç. E., Gunindi Z., Habibi F., Yazıcı Z. A., Ekenel H. K. VIDI: A Video Dataset of Incidents. *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP 2022).*

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- **Sesver D.**, Tuna S., Aktaş M. S., Kalıpsız O., Kanlı A. N., Turgut U. O. (2019). Implementation of Association Rule Mining Algorithms on Distributed Data Processing Platforms. *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 79-84, 2019.
- **Sesver D.**, Küçük S. E., Aras A. C., Kazimi M., Zengin E., Dohr M. Battery Age Estimation: A Data Driven Approach, *International Conference & Exhibition on DIGITAL TRANSFORMATION SMART SYSTEMS (DTSS)* October 2019, page 47-50, Ankara, Turkey.