ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

RADAR TARGET DETECTION USING IMPROVED TRANSFORMER NEURAL NETWORKS

M.Sc. THESIS

Sena ÇAYBAŞI

Department of Electronics and Communication Engineering

Electronics Engineering Programme

JUNE 2025

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

RADAR TARGET DETECTION USING IMPROVED TRANSFORMER NEURAL NETWORKS

M.Sc. THESIS

Sena ÇAYBAŞI
(504211253)

Department of Electronics and Communication Engineering

Electronics Engineering Programme

Thesis Advisor: Prof. Dr. Işın ERER

JUNE 2025

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

GELİŞTİRİLMİŞ TRANSFORMER SİNİR AĞLARI İLE RADAR HEDEF TESPİTİ

YÜKSEK LİSANS TEZİ

Sena ÇAYBAŞI
(504211253)

Elektronik ve Haberleşme Mühendisliği Anabilim Dalı

Elektronik Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Işın ERER

HAZİRAN 2025

Sena ÇAYBAŞI, a M.Sc. student of ITU Graduate School student ID 504211253 successfully defended the thesis entitled "RADAR TARGET DETECTION USING IMPROVED TRANSFORMER NEURAL NETWORKS", which he/she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Işın ERER** Istanbul Technical University	.....
Jury Members :	**Prof. Dr. Mesut KARTAL** Istanbul Technical University	.....
	Prof. Dr. Selçuk PAKER Istanbul Nişantaşı University	.....

Date of Submission : **30 May 2025**
Date of Defense : **23 June 2025**

To my family and friends,

FOREWORD

I would like to express my endless thanks to my family for their support throughout my entire life, including my education life. I would like to represent thanks to my friend Begüm Cangöz, who supported me during my thesis. I would like to express my deepest gratitude thanks my advisor Prof. Dr. Işın Erer for her academic support and encouragement throughout my graduate education. I would like to express my special thanks to my managers, ASELSAN Inc and Aselsan Academy, of which I am a member, for supporting me during my graduate education.

June 2025 Sena ÇAYBAŞI

TABLE OF CONTENTS

ABBREVIATIONS

CNN : Convolutional Neural Network
ViT : Vision Transformer
CFAR : Constant False Alarm Rate
SNR : Signal to Noise Ratio
RCS : Radar Cross Section
CW : Continuous Wave
FMCW : Frequency-Modulated Continuous Wave
MTI : Moving Target Indication
CA : Cell Averaging
SOCA : Smallest Of Cell Averaging
GOCA : Greatest of Cell Averaging
OS : Ordered Statistics
RNN : Recurrent Neutral Network
NLP : Natural Language Processing
LSTM : Long Short-Term Memory
SAR : Synthetic Aperture Radar
IF : Intermediate Frequency
RF : Radio Frequency
CUT : Cell Under Test
MLP : Multi-Layer Perceptron
ReLU : Rectifier Linear Unit
LO : Local Oscillator
ADC : Analog to Digital Converter
STC : Sensitivity Time Control
FIR : Finite Impulse Response
NLFM : Non-Linear Frequency Modulation
LFM : Linear Frequency Modulation
MTI : Moving Target Indicator
PRF : Pulse Repetition Frequency
PRI : Pulse Repetition Interval
FFT : Fourier Transform
PRI : Pulse Repetition Interval
FFN : Feed-Forward Neural Network
CTF-Net : Convolutional and Transformer Fusion Network
LN : Layer Normalization
W-MSA : Window-Based Multi-head Self Attention
SW-MSA : Shifted Window-Based Multi-head Self Attention
ADAM : Adaptive Moment Estimation

SYMBOLS

t	**:** Time
P_r	**:** Received power
P_t	**:** Transmitted power
G_t	**:** Tranmitter antenna gain
G_r	**:** Receiver antenna gain
λ	**:** Wavelength
f	**:** Frequency
R	**:** Range
L	**:** System losses
Q	**:** Quadrature component of the signal
I	**:** In-phase component of the signal
Q	**:** Query matrix
K	**:** Key matrix
V	**:** Value matrix
W_i^Q**,**	**:** Transformation of Query, Key, Value
W_i^K	**:** Transformation of Key
W_i^V	**:** Transformation of Value
W_i^O	**:** Output mapping
d_m	**:** the size of the encoding vector
$d_m x d_q$	**:** Dimensions of query
$d_m x d_k$	**:** Dimensions of key
$d_m x d_v$	**:** Dimensions of value
P_{fa}	**:** Probability of false alarm
P_d	**:** Probability of detection

LIST OF TABLES

Page

LIST OF FIGURES

Page

RADAR TARGET DETECTION
USING IMPROVED TRANSFORMER NEURAL NETWORKS

SUMMARY

Radar systems, which use electromagnetic waves to determine a target's position, speed, and direction, are essential in a variety of fields, including mapping, weather forecasting, and surveillance. Despite having different areas of use, the most basic function of radar systems is target detection. Target detection in radars is defined as the process of processing raw radar data and extracting meaningful information.

In the process of extracting meaningful information, data obtained from the real world is usually complex and contains clutter. Clutter in the environment can make it difficult to understand the data received from the radar. Target detection can be challenging due to interferences, multiple reflections, low radar cross-section areas of targets and environmental factors. Therefore, it is of critical importance that radar signal processing methods operate with high accuracy and reliability, especially during the target detection phase.

In traditional methods; in order to perform radar target detection, a fixed threshold level is determined to remain above the level of the signal reflected from the environment. The target detection decision is made by examining the signals remaining above this level. However, in cases where the noise and environment distribution are variable, applying a fixed threshold causes false detections. In order to prevent this situation, an adaptive threshold application suitable for changing conditions is required.

One of the widely used adaptive thresholding methods is the Constant False Alarm Rate (CFAR) algorithms. CFAR determines the threshold value based on the statistical properties of the reference cells located around the target cell and perceives signals above this threshold value as targets. The main advantage of this algorithm is that it can keep the false alarm rate under control.

Due to their reliance on predetermined statistical models and susceptibility to environmental changes, traditional approaches like CFAR algorithms frequently struggle with target detection in complex environmental situations. To get over these restrictions, deep learning-based approaches have started to be employed as alternatives to conventional methods.

One of the first methods encountered in the literature on target detection with deep learning is CNN. Convolutional Neural Networks (CNN) are capable of learning and identifying characteristics in images. It is anticipated that they have a great deal of potential for radar target detection because of this property. However, CNNs generally focus on pixels in a small area, making it difficult for them to capture connections between larger and more distant areas in the image. As a result, understanding of the overall context is limited. To overcome this issue, transformer-based models such as

Vision Transformer (ViT) divide the image into small pieces, analyze them sequentially, and use an attention mechanism to learn the long-distance relationships between these pieces. However, due to the complex structure of ViT, Swin Transformer models appear in the literature as an alternative. Swin Transformer models, which implement an attention mechanism through windowing, are more effective models in terms of complexity. Although Swin Transformer appears to be more efficient in terms of model complexity, the application of an attention mechanism through windowing can cause the model to learn insufficiently and lead to imbalances during training in data where global relationships are important, such as radar data. In addition, since Swin Transformer is a model that performs effectively, especially with large data sets, it may not generalize sufficiently when the data set is insufficient, making it more prone to overfitting. This situation can lead to imbalances during the training process.

Since ViT has a more complex structure than the CNN architecture, it requires larger data sets to work with. Although Swin Transformer models are less complex than ViT due to their requirement for larger data sets and lack of data generalization, it is not possible to achieve higher target detection performance with less data. Instead, there is a need for models that can better learn global relationships by considering both the range and Doppler axes in radar data and that have higher generalization capabilities. For this reason, due to insufficient data source, this thesis proposes new radar target detection methods based on an improved ViT architecture.

In this study, two different approaches are suggested. First, an improved feedforward network structure is used instead of the Multi-Layer Perceptron (MLP) in standard ViT architectures. This model is called improved ViT. This development aims to increase the learning capacity of the model. It especially aims to improve generalization in complex environments where clutter and noise are intense. Second, a hybrid model is designed by integrating the improved ViT architecture with a CNN in parallel. In hybrid model, the feature maps obtained from the input image are sent to both CNN and improved ViT in parallel. Both structures process features in a way that is specific to their architecture. While CNNs provide superiority in extracting local features from visual data, ViT has the ability to model long time dependencies. A rich representation containing both local and global information is created at the end layer by concatenating the feature maps from the two pathways. The hybrid model created by combining these two models aims to use the strengths of both architectures.

Radar data is two-dimensional data consisting of Range and Doppler axis. If these are interpreted as snapshots of the environment, image-based deep learning models such as CNN and ViT can be used for radar data. The proposed approaches are tested on both synthetic and real radar datasets for comparison.

The suggested models have been evaluated empirically by contrasting them with CNN-based method and traditional CFAR algorithms such as CA-CFAR, SOCA-CFAR, GOCA-CFAR, and OS-CFAR. The results consistently demonstrate that the hybrid CNN-ViT and improved ViT models perform better than the conventional techniques in terms of detection accuracy, particularly in environments with clutter. This clearly demonstrates that the integration of deep learning architectures into the radar signal processing pipeline leads to better detection performance. This study offers convincing proof that deep learning-based techniques, with their increased flexibility, adaptive

learning capabilities, and improved feature representation, can outperform conventional radar detection methods.

GELİŞTİRİLMİŞ TRANSFORMER SİNİR AĞLARI İLE RADAR HEDEF TESPİTİ

ÖZET

Radar sistemleri; elektromanyetik dalgalar vasıtasıyla hedeflerin konum, hız ve yön bilgilerini tespit eden, gözetleme, haritalama, hava durumu bilgisi üretme gibi birçok farklı alanda önemli bir yere sahip olan sistemlerdir. Farklı kullanım alanlarına sahip olmasına rağmen radar sistemlerinin en temel fonksiyonu hedef tespitidir. Radarlarda hedef tespit işlemi, ham radar verilerinin işlenerek anlamlı bilgiler çıkarma süreci olarak tanımlanır.

Anlamlı bilgiler çıkarma sürecinde, gerçek dünyadan elde edilen veriler genellikle karmaşık yapıdadırlar ve gürültü barındırırlar. Ortamın kargaşa (clutter) barındırması, radardan alınan verilerin anlaşılmasını zorlaştırabilmektedir. Girişimler, çoklu yansımalar, hedeflerin düşük radar kesit alanlarına sahip olması ve çevresel faktörler nedeniyle hedef tespiti zorlu hale gelebilmektedir. Bu nedenle radar sinyal işleme yöntemlerinin; özellikle hedef tespiti aşamasında, yüksek doğruluk ve güvenilirlikte çalışması kritik bir öneme sahip olmaktadır.

Radar verileri, genellikle Range-Doppler (menzil-Doppler) matrisi biçiminde temsil edilirler. Bu matrisin yatay ekseni, hedefin Doppler frekans kaymasını; yani hedefin hareket hızını gösterirken, dikey eksen ise hedefin menzil bilgisini içermektedir. Range-Doppler matrisi, ortamın iki boyutlu bir görüntüsü olarak ele alınabilmekte ve görüntü işleme teknikleri kullanılarak analiz edilebilmektedir.

Geleneksel yöntemlerde; radar hedef tespiti yapabilmek için; ortamdan yansıyan sinyal seviyesinin üstünde kalacak şekilde sabit bir eşik seviyesi belirlenmektedir. Bu seviyenin üzerinde kalan sinyallere bakılarak, ilgili verinin tespit olduğuna dair yorum yapılabilmektedir. Ancak gürültünün ve ortam dağılımının değişken olduğu durumlarda, sabit eşik uygulanması yanlış tespitlerin ortaya çıkmasına sebep olmaktadır. Bu durumun önüne geçmek için, değişen koşullara uygun adaptif bir eşik uygulaması gerekmektedir.

Yaygın olarak kullanılan adaptif eşikleme yöntemlerinden biri, Sabit Yanlış Alarm Oranı (Constant False Alarm Rate, CFAR) algoritmalarıdır. CFAR, hedef hücresinin çevresinde yer alan referans hücrelerinin istatistiksel özelliklerini baz alarak eşik değer belirlemekte ve bu eşik değerin üzerindeki sinyalleri hedef olarak algılamaktadır. Bu algoritmanın temel avantajı, yanlış alarm oranını kontrol altında tutabilmesidir.

Hücre Ortalama (Cell Averaging, CA), En Küçük Hücre Ortalama (Smallest Of Cell Averaging, SOCA), En Büyük Hücre Ortalama (Greatest of CFAR, GOCA) ve Sıralı İstatistikler (Ordered Statistics, OS) en yaygın kullanılan CFAR algoritmalarındandır. CA-CFAR, çevredeki referans hücrelerin ortalama güç seviyesini hesaplayarak tespit

eşiğini belirlemektedir. SOCA-CFAR, en düşük güç seviyesine sahip referans hücreleri kullanarak adaptif eşik belirlemekte ve özellikle düşük SNR ortamlarında yanlış alarmları azaltmada etkili bir yöntem olarak literatürde geçmektedir. GOCA-CFAR ise en yüksek güç seviyesine sahip referans hücrelere dayalı bir eşik belirleyerek, yoğun hedef bölgelerinde daha iyi tespit performansı sağlamaktadır. OS-CFAR, referans hücreler içindeki sıralı istatistiklere dayanarak eşik belirleme yapmakta ve heterojen kargaşa ortamlarında daha dengeli bir tespit performansı sunmaktadır. Genellikle kargaşadan etkilenilen ortamlarda OS-CFAR tercih edilmesiyle beraber bu yöntemler, radar ortamındaki gürültü ve kargaşaya adapte olarak yanlış alarm oranını kontrol altında tutmayı amaçlamaktadır. Ancak CFAR yöntemlerinin başarısı, radar ortamının homojen ve istatistiksel varsayımlara uygun olmasına bağlı olması nedeniyle CFAR tabanlı geleneksel yöntemler gerçek radar uygulamalarında özellikle yüksek karmaşıklık ve gürültü altında sınırlamalarla karşılaşmaktadır.

CFAR algoritmaları gibi geleneksel yaklaşımlar önceden belirlenmiş istatistiksel modellere dayalı olmaları ve çevresel değişikliklere karşı duyarlı olmaları nedeniyle hedef tespiti zor bir problem haline gelmektedir. Bu kısıtlamalardan kurtulmak için, geleneksel yöntemlere alternatif olarak derin öğrenme tabanlı yaklaşımlar kullanılmaya başlanmıştır.

Derin öğrenmeyle hedef tespiti konusunda literatürde karşımıza çıkan ilk yöntemlerden biri CNN'dir. Evrişimli Sinir Ağları (Convolutional Neural Network, CNN), görüntülerdeki özellikleri öğrenme ve tanımlama yeteneğine sahiptirler. Bu özellikleri nedeniyle, radar hedef algılaması için büyük bir potansiyele sahip oldukları görülmektedir. Ancak CNN'ler genellikle görüntüde yer alan küçük bir alandaki piksellere odaklanmaktadır ve bu da görüntüdeki daha büyük ve daha uzak alanlar arasındaki bağlantıları yakalamalarını zorlaştırmaktadır. Sonuç olarak, görüntüdeki genel bağlamın çıkarılması CNN için zor bir problem haline gelmektedir. Bu sorunu aşmak için, Görsel Dönüştürücü (Vision Transformer, ViT) gibi transformer tabanlı modeller tercih edillmektedir. ViT görüntüyü küçük parçalara ayırıp analiz etmektedir ve bu parçalar arasındaki uzun mesafeli ilişkileri öğrenmek için bir dikkat mekanizması kullanmaktadır. Bu mekanizma sayesinde, her bir girdinin bağlam içirisindeki önemi değerlendirilmekte ve uzun vadeli ilişkiler yakalanarak bilgiyi daha etkili modellemesine imkan tanımaktadır. Ancak ViT'in sahip olduğu karmaşık yapı sebebiyle buna alternatif olarak literatürde karşımıza Swin Transformer modelleri çıkmaktadır. Pencereleme üzerinden dikkat mekanizması gerçekleyen Swin Transformer modelleri karmaşıklık bakımından daha efektif modellerdir. Her ne kadar Swin Transformer, model karmaşıklığı bakımından daha verimli görünse de, pencereleme üzerinden dikkat mekanizması uygulandığı için radar gibi küresel ilişkilerin önemli olduğu verilerde modelin yetersiz öğrenmesine ve eğitim sırasında dengesizlikler olmasına neden olabilmektedir. Ayrıca Swin Transformer, özellikle büyük veri kümelerinde etkin performans sergileyen bir model olduğundan, veri kümesinin yetersiz olduğu durumlarda yeterli genelleme yapamayarak aşırı öğrenmeye daha yatkın hale gelmektedir. Bu durum eğitim sürecinde dengesizliklere yol açabilmektedir.

ViT, CNN mimarisinden daha karmaşık bir yapıya sahip olduğundan, daha büyük veri kümeleriyle çalışması gerekmektedir. Swin Transformer modelleri ise ViT'e göre daha büyük veri kümeleri gerektirmesi ve veri genellemesindeki eksikliğinden dolayı daha

az karmaşıklığa sahip olsa da az veri ile daha yüksek bir hedef tespit performansı elde etmek mümkün olmamaktadır. Bunun yerine, radar verisinde menzil ve Doppler eksenlerinin birlikte ele alınıp global ilişkilerin daha iyi öğrenilebileceği ve genelleme kabiliyeti daha yüksek olan modellere ihtiyaç duyulmaktadır. Bu sebeple, bu tez, az veri ile daha yüksek performans göstermesini sağlamak için geliştirilmiş ViT mimarisine dayalı yeni radar hedef tespit yöntemleri önermektedir.

Bu tez çalışmasında, modifiye edilmiş iki farklı derin öğrenme tabanlı model kullanılmaktadır. Standart ViT yapılarında yer alan Multi-Layer Perceptron (MLP) katmanı, özellikle karmaşık radar ortamlarında modelin genelleme kapasitesini sınırlandırabildiği için ilk olarak MLP katmanı yerine, evrişim katmanı içeren geliştirilmiş bir ileri ağ (feedforward) yapısı önerilmektedir. Bu yapı ile hedef tespitinin kargaşa varlığında, daha doğru bir şekilde yapılması amaçlanmaktadır.

Önerilen diğer yöntem ise, sinyallerin sahip olduğu özniteliklerin daha etkili bir şekilde çıkarılması ve işlenmesine olanak tanıyan hibrit bir modeldir. Bu model, paralel olarak çalışan bir CNN ve geliştirilmiş ViT mimarisinden oluşmaktadır. Hibrit modelde, giriş görüntüsünden elde edilen özellik haritaları hem CNN hem de geliştirilmiş ViT yapısına paralel olarak gönderilmektedir. Her iki yapı da özellikleri kendi mimarilerine özgü bir şekilde işlemektedir. CNN yapısı; özellikle görüntü verisindeki kenar, doku ve şekilleri başarılı şekilde çıkarırken; ViT'ler uzun süreli bağıntıları modellemede üstünlük sağlamaktadır. Bu iki yapının kombinasyonu, hedef tespit performansını artırıcı etki meydana getirmektedir.

Radar verisi, ortama bağlı olarak karmaşık ve yüksek kargaşa içeren veriler barındırabilmektedir. Bu nedenle, derin öğrenme modellerinin eğitim aşamasında, modelin hem çevresel hem de zamansal özellikleri öğrenebilmesi model başarısı için etkin rol oynamaktadır.

Bu çalışmada kullanılan veri seti, hem sentetik hem de gerçek radar sistemlerinden toplanmış veriler ile oluşturulmuştur. Sentetik veriler, hedeflerin farklı konum, hız ve genlik özelliklerine göre çeşitlendirilirken; gerçek veriler ise gürültü, kargaşa gibi çevresel etmenleri içeren sinyaller barındırmaktadır.

Model eğitimi sürecinde, sentetik ve gerçek veri kümesi eğitim, doğrulama ve test alt kümelerine ayrılmış; benzetimler sentetik ve gerçek veriler için ayrı ayrı gerçekleştirilmiştir. Elde edilen bulgular; önerilen ViT ve CNN-ViT hibrit modellerinin, geleneksel CFAR ve CNN modellerine kıyasla daha yüksek hedef tespit performansı sağladığını göstermektedir. Sadece gürültünün bulunduğu sentetik verilerde, tüm derin öğrenme yöntemlerinin hedef tespit doğruluk oranı %98 mertebelerinde kalırken; gerçek radar ortam koşullarında CNN yönteminin %88, önerilen ViT modelinin %89 ve önerilen hibrit modelin %90 mertebelerinde hedef tespit doğruluğu sağladığı gözlemlenmiştir.

Sonuç olarak bu tez çalışmasında, geleneksel radar hedef tespit algoritmalarına alternatif yöntemler önerilmiştir. Önerilen yöntemler, klasik yöntemlerin ötesine geçen, derin öğrenme tabanlı yeni yaklaşımlar sunmaktadır. Geliştirilmiş ViT mimarisi ve CNN-ViT hibrit modeli, radar hedef tespitinde doğruluk ve karmaşık ortam koşullarına adaptasyon açısından önemli avantajlar sağlamaktadır.

1. INTRODUCTION

Radar systems are technologies that detect the position, speed, and direction of targets through electromagnetic waves and have a crucial place in many different areas, such as surveillance, mapping, and weather forecasting. Its acronym is "radar detection and ranging". The history of the concept of radar dates back to the early 1900s, with the first explanation of electromagnetic waves and their propagation. Although several countries were testing radio wave detection in the 1930s, the period when the development of the radar concept accelerated dates back to the years of World War II. Initially designed for military purposes, radar's utility beyond conflict quickly became clear. Subsequent to the war, technology rapidly integrated into daily life. Law enforcement departments, for example, monitor vehicle speeds worldwide and enforce traffic rules using radar. In-depth atmospheric research and timely weather reporting are produced by meteorologists using radar. Radar is absolutely essential in aviation for tracking aircraft height, monitoring air traffic, and helping pilots avoid hazardous weather. To improve runway visibility in inclement weather, new radar technologies are being developed. Radar helps boats on the water navigate safely and avoid collisions by detecting buoys and other indicators. Road vehicle safety also benefits from similar technology. Additionally, satellites and aerial radar have evolved into important instruments for recording the Earth's surface and tracking ecological shifts over time [1]. Over other detection systems, radar systems have numerous advantages. They are able to operate efficiently throughout the day because they are not reliant on light, which is the primary advantage. Additionally, they are considerably less vulnerable to adverse weather conditions, such as fog, rain, snow, and grit, as a result of their employment of electromagnetic radiation. Radars can therefore continue to produce reliable data in situations where the efficacy of optical or infrared systems is significantly reduced. On the other hand, radars are a substantial advantage, particularly for long-range surveillance applications and early warning systems, as they are capable of detecting signals at an extremely high capacity over extremely long distances [2]

In general, radars operate by transmitting a signal with a specific waveform into the environment and detecting an echo signal that is reflected from the environment. Basically, Radar system consists of a receiving antenna, a receiver, and a transmitting antenna that emits electromagnetic radiation produced by an oscillator. When a radar signal reaches a target, some of its transmitted energy is reflected and scattered in various directions. What is important is how much of this reflected energy reaches the radar. The reflected signal from the target reaches the radar receiver antenna and is then transmitted to the receiver level. The signal arriving at the receiver is interpreted to determine the presence, position and speed of the target. The target's range is calculated from the information on how long it takes for the signal sent from the radar to reach the radar again. The target's direction, or angular position, can be ascertained from the arrival direction of the reflected wavefront [3].

While the basic signal structures and operating principles of radars may differ, radars ultimately serve the same critical function: to detect and localize targets. Target detection, a fundamental task of radar systems, involves determining the presence of objects of clutter, noise, and other environmental interference. Radar signal detection must be regarded as a statistical problem due to the presence of interference and noise [4]. Echoes from the ground, or clutter, can occasionally be the target and other times they are interference. Ground clutter, noise, and possibly jamming are the interference while attempting to detect a moving vehicle; but, when imaging a specific area of the earth, the same topography becomes the intended target, and the only interference is noise and jamming [1]. However, for a surveillance radar clutter can be mentioned as the unwanted echoes in the system that are generated by objects such as terrain, structures, or weather phenomena.

Interference is the result of various electromagnetic sources that can interfere with the radar signal, whereas noise is the random background or thermal fluctuations that exist in the radar receiver. These things can reduce the radar target detection performance by masking the target or distorting the signal and it causes lower The Signal-to-Noise Ratio (SNR). SNR is a primary metric in radar systems that quantifies the intensity of the desired signal in relation to the background noise. This directly enhances the reliability of target detection. Higher SNR determines that the signal is more distinct

from noise. For detecting weak and distant targets, it's crucial that achieving an adequate SNR. Higher SNR is needed for accurate target detection, especially in environments with high interference and chaos. There are some factors that affects the SNR such as transmitted power, antenna gain, propagation losses, and receiver sensitivity [5]. To prevent loss in SNR, radar systems apply various strategies, including adaptive thresholding and signal processing algorithms such as Doppler filtering or Moving Target Indication (MTI). In addition to system design improvements, the reliability and accuracy of detection can be improved by applying interference suppression methods, frequency hopping and efficient radar detection algorithms.

1.1 Purpose of Thesis

Under changing environmental conditions, radar detection becomes a very difficult problem for targets with low SNR. Target detection, which is the basic task of radars, is usually done by comparing the signal coming to the receiving line with a threshold value in traditional methods. When the signal strength exceeds the specified threshold value, the target presence is detected. Dynamic environmental conditions have led to the need to determine an adaptive threshold value instead of a fixed threshold for target detection. Constant False Alarm Rate (CFAR) algorithms have been introduced to the literature to serve this purpose and are among the algorithms that use adaptive threshold values [6]. However, in situations where the clutter is intense and target detection becomes harder. False detections due to chaos also occur even though there is no real target. In order to increase target detection performance in environments with intense clutter, deep learning-based methods are needed in addition to traditional methods. Deep learning based methods such as CNN, Vision Transformer (ViT), Swin Transformer are used for deep learning radar applications in the past. In this study, an alternative deep learning based method is proposed in addition to traditional CFAR methods, CNN method and ViT based methods in radar target detection, which is one of the fundamental problems of radar systems. In the study, both range-Doppler dataset was generated using radar simulator and real field data was studied. The proposed methods are compared with

different CFAR algorithms, CNN method, ViT and Swin Transformer methods in terms of target detection performance.

1.2 Literature Review

CFAR is a method that provides ease of use in finding targets by determining the noise level in the environment. Cell Averaging (CA), Smallest Of Cell Averaging (SOCA), Greatest of CFAR (GOCA) and Ordered Statistics (OS) are the most widely used CFAR algorithms. However, depending on variable environmental conditions; especially in environments containing clutter, it becomes quite difficult to distinguish between target signals and clutter, which reduces the success of traditional methods [7].

Deep learning-based approaches, which have gained rapid development momentum in recent years, have begun to be used as an alternative to traditional methods. In image classification, convolutional neural networks have achieved remarkable results [8] [4]. Furthermore, natural language processing has made effective use of deep learning [9]. To cope with sequential data, such as audio signals and words, researchers employ recurrent neutral networks (RNNs) to do convolutional decoding [10], speech recognition [11], text classification [12], and other tasks.

These days, radar researchers have attempted to use deep learning techniques to accomplish radar related tasks. Prior research on artificial intelligence-based radar applications frequently emphasized the significance of man-made features. Haykin employed neural networks to tackle the radar clutter classification problem since they are naturally able to extract features from input data [13]. As a result of the increasing success of networks in image-related tasks, researchers have used CNNs to deal with radar signals by transforming signals into time-frequency representation images for efficient target recognition [14]. Even in complex and non-stationary cluttered situations, these CNN-based techniques have shown encouraging results in recognizing radar targets [15]. However, work by the authors Yuan Xie, Jun Tang, and Li Wang has also demonstrated an alternative method that goes over the conventional use of time-frequency analysis images [16]. They proposed to feed the CNN architecture directly with range-Doppler spectrum data instead of time-frequency images. This technique aims to improve target detection performance in high complexity

environments by exploiting the spatial properties of range-Doppler data and offers a new perspective on deep learning applications in radar target recognition. CNN have the ability to learn and recognize features in images. Thanks to this feature, it has been predicted that they will also have significant potential in radar target detection [17]. As a result of this prediction, it can be seen in many studies in the literature that CNN-based methods have higher accuracy rates compared to traditional methods [7,17].

Later, as the popularity of convolutional neural networks grew, researchers used them to build signal detectors based on combined time-frequency analysis images [14] and to identify radar targets in highly complex and non-stationary distributed scenes [15]. To realize target detection in a novel way, the authors of the [16] study feed the range-Doppler spectrum to the CNN instead of using time-frequency analysis images as input.

Since radar echo signals are temporally dependent in nature, much research on radar-based signal processing utilizes long short-term memory (LSTM) networks or hybrid CNN-LSTM models. The sequential nature of such data is best modeled by LSTM networks that capture the relationship between past and current inputs [18]. These networks can improve accuracy and reduce model complexity by using the temporal features of the radar. Ma et al. proposed a fusion model in which LSTMs describe temporal dynamics and CNNs extract spatial features from Doppler maps [19]. The introduction of attention-based mechanisms such as self-attention further improved feature representation. These allow the model to facilitate parallel data processing and improve temporal-spatial feature integration by computing the similarity between any two regions in a micro-Doppler radar image.

Transformer neural networks are mostly used in the field of natural language processing (NLP), but are also frequently used in image classification, object detection, and segmentation. In recent years, the use of transformer-based models in radar signal processing has also begun to take place in the literature. In place of the CNN structure, it's presented in [20] that the Transformer structure [21]. Detecting changes in remote sensing images over time is considered as challenging problem. The main reason for this is that the same objects may appear differently at different times, and traditional

methods cannot distinguish these differences well. CNN-based approaches are also inadequate in capturing distant relationships between images. To solve this problem, a transformer based model is proposed in [22]. This model extracts features from images at two different times, simplifies this information into more meaningful representation, and learns the relationships between this information with the transformer structure. Then, it converts this information back to the image level and produces a change map. It's expressed that the method used less computational power and outperformed existing methods by working with high accuracy on many data sets.

In autonomous driving fields, in order to detect driver distraction cases, an approach is presented where a transformer model with stronger relationship capturing ability compared to CNN/RNN based methods is applied. This model uses an architecture that both captures long-term spatial relationships and offers computational efficiency by limiting attention calculations with a special phased structure and window shifting mechanism [23]. In areas such as image captioning [24], transformers, especially the Vision Transformer (ViT) [25], have attracted attention for their impressive performance in image understanding tasks. However, situations with high-resolution images or a small amount of training data are not suitable for ViT due to its computational requirements. To overcome these issues, the Swin Transformer [26] is presented as a hierarchical Transformer with shifted windows that preserves performance while reducing the computational load. However, the relatively large model size restricts deployment on peripheral devices with limited processing and memory capacity. As a solution to these problems, Gu and colleagues developed IR-ST, a lightweight neural network that uses FMCW radar data to detect human falls [27]. The Swin-Transformer backbone is integrated with an inverted residual module in IR-ST to achieve a balance between detection sensitivity and model efficiency. As a potential method for radar-based human activity recognition, this architecture is particularly well suited for real-time applications on resource-limited hardware.

Transformer designs incorporate an attention mechanism that allows them to effectively eliminate irrelevant background clutter while enhancing important target features. Recognizing this benefit, Wu et al. presented CTF-Net, a novel architecture that combines Transformers and CNNs for synthetic aperture radar (SAR) ship detection.

The ability to combine local and global feature extraction capabilities has attracted attention in recent studies for hybrid architectures fusing Transformers and CNNs. Using this design, CTF-Net successfully handles the challenges offered by complex marine environment, including low target contrast, intense background mess and small item sizes. The network can catch both wide spatial relationships and detailed textures, which eventually enhances detect accuracy and stability, which is thanks to integration of convolutional and Transformer modules. For SAR-based object detection works where traditional models will show poor performance, this CTF-Net makes it an effective option.

2. RADAR DETECTION

2.1 Radar Data

Radar equation is the most important parameter to understand the performance of the radar and optimize the system design. The physical relationships of the energy transmitted by the radar from the propagation of the wave to the reception of the reflected signal are explained by the radar equation. It calculates how much power the radar signal loses from the time it reaches the target to the time it returns. The transmitted power (P_t) shows how strong a wave the radar emits. The gain of the antennas that transmitter antenna Gain (G_t) and receiver antenna gain (G_r), determines how effectively that signal is directed and collected. The wavelength of the signal (λ) depends on the frequency and it affects the reflection of the signal. Radar Cross Section (RCS), σ, is a measure of how much a target reflects the signal from the radar. For instance, it can be detected small targets using short wavelength radar. RCS depends on the physical properties of the target, such as its shape, size, orientation, and material. As the range of the target (R) increases, the strength of the signal decreases significantly because the signal both travels and returns. Finally, losses (L) in the system or environmental factors can also reduce the signal power. When all these variables taken together, it indicates how much power reaches the radar receiver and how far away the radar can detect a target [28]. The radar equation is showed in equation 2.1:

$$P_r = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 R^4 L} \tag{2.1}$$

A diagram of a typical radar system is shown in Figure 2.1 [29].

Radar systems are categorized into two groups according to the signal type: Pulse and Continuous Wave (CW) radars. Pulse radars, which can send high-power signals, determine the target's location and distance by measuring the time it takes to hit and return from the target [3]. Due to their long-range and precise target detection

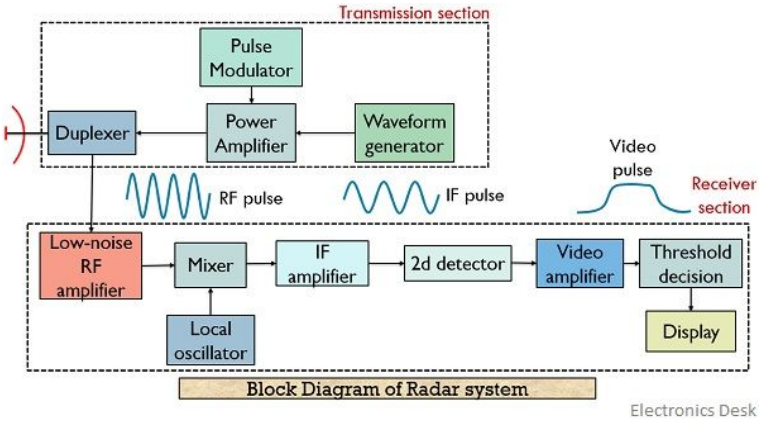


Figure 2.1: A standard radar block diagram.

capacity, they are preferred in early warning, air defense and surveillance systems where the timely detection of high-altitude and high-speed moving air targets is of critical importance. In addition, since pulsed radars have a wide scanning area, they can track multiple targets simultaneously, which increases air defense effectiveness in multi-threat scenarios. CWradars operate by transmitting uninterrupted signals and determine a target's velocity through the Doppler effect [30]. However, they can't determine the target's range. Frequency-Modulated Continuous Wave (FMCW) radars use signals whose frequency changes over time to overcome this restriction. So, by using FMCW radar it can be extracted both speed and distance. FMCW radars are widely used in level measurement sensors, automotive systems, and other short-range detection applications because of this dual capability [1]. Different radar waveforms are shown in Figure 2.2 [31].

The radar data set used in this study is based on a pulse-Doppler radar. In pulse-Doppler radars, signals reflected from the environment are received by the radar receiver. The collected data is first sampled in the fast-time range axis and converted into digital form. After the IF (Intermediate Frequency) or RF (Radio Frequency) signal is downsampled to baseband, it goes through noise filtering and matched filtering. After all pulses have been collected and fast-time processing is completed, slow-time inter-pulse processing is started for coherent radars. Depending on the application, Moving Target Indicator (MTI), Doppler filtering and windowing can be applied in inter-pulse processing. For the data used in this study, Doppler filtering was applied in inter-pulse processing. The data

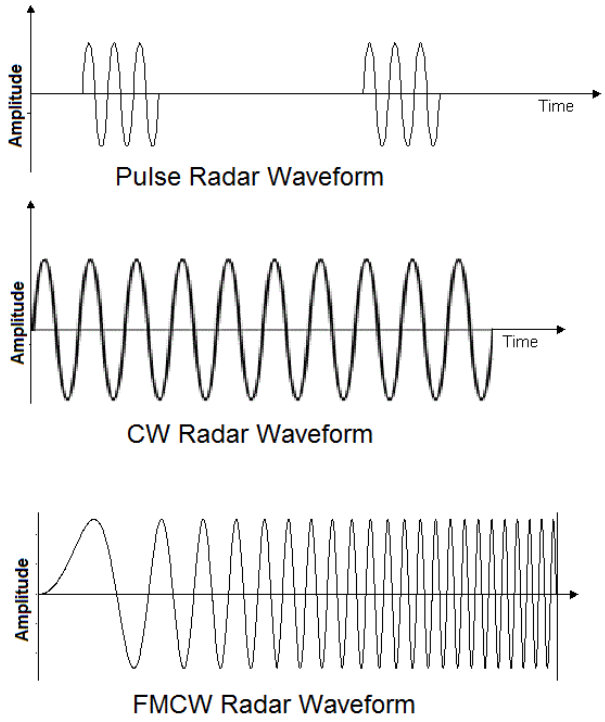


Figure 2.2: Radar waveforms.

obtained as a result of fast-time and slow-time processing is called the range-Doppler matrix. Since detection algorithms are power-based algorithms, all operations are performed on the square amplitude and the input of the detection algorithms is the absolute square of the range-Doppler matrix amplitude values. Figure 2.3 shows an example of radar data.

The process of radar signal processing steps are explained in the following sections.

2.2 Radar Signal Processing

Radar signal processing consists of fast-time processing, which covers intra-pulse operations, and slow-time processing, which covers inter-pulse operations. Following sections will describe the fast and slow time processings.

2.2.1 Fast-time processing

Fast-time signal processing includes time-dependent operations within a pulse duration. These are sampling, baseband downconversion, noise filtering and matched filtering.

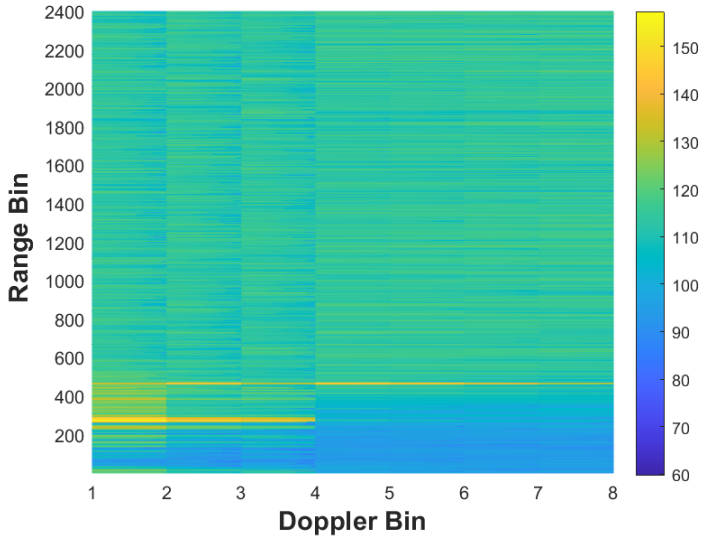


Figure 2.3: Range-Doppler matrix data.

2.2.1.1 Sampling

In radars, the data received from the environment must be digitized in order to be processed. Therefore, the first step in radar signal processing is sampling. Sampling is the process of taking samples from data over time. The sampling process is performed with the help of Analog to Digital Converters (ADC). The sampling frequency of the ADC refers to the number of times per second the received analog signal is sampled. According to sampling Nyquist's Theorem, the sampling rate should be at least twice as fast as the highest frequency component of the received signal. Sampling can be done directly on an RF signal or at the IF level by downsampling the signal to an intermediate frequency. There is no need for mixer structures and LO as no frequency conversion is required during direct sampling. RF signal can be directly sampled at high speed. Since it is sampled in RF, there is no need for IF circuit elements and direct sampling systems have a smaller hardware footprint. It's demonstrated RF sampling and IF sampling diagram in figure 2.12 [32].

For radars to process the reflected signal, the incoming signal must be within the dynamic range of the ADC. The dynamic range is the difference between the strongest and weakest signals from the ADC. It is related to the bit resolution of the ADC and is usually expressed in decibels (dB). Radar systems need a wide dynamic range

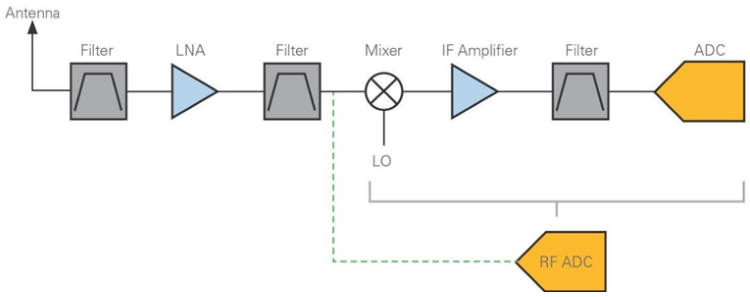


Figure 2.4: RF and IF sampling.

because target distance, radar cross section and ambient factors can all affect how strongly signals are reflected. If the dynamic range is too small, either strong signals are saturated and distorted or weak signals cannot be recognized. However, it may not always be possible to have a wide dynamic range due to hardware limitations. For this reason, some filtering is needed to keep the reflected signal within the dynamic range of the ADC. These can be limiters or Sensitivity Time Control (STC). Limiter is used to limit the incoming signal while STC applies attenuation to the signal depending on the time. This ensures that the incoming signal remains within the dynamic range of the ADC. STCs are used to suppress high amplitude signals from close by and pass low amplitude signals from distant targets. In this way, ADC saturation is prevented. It's demonstrated an example of STC filter in figure 2.13 [33].

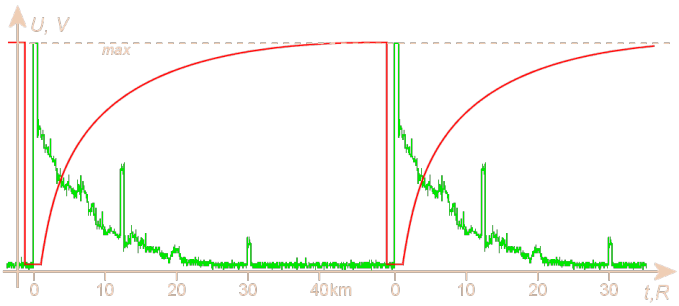
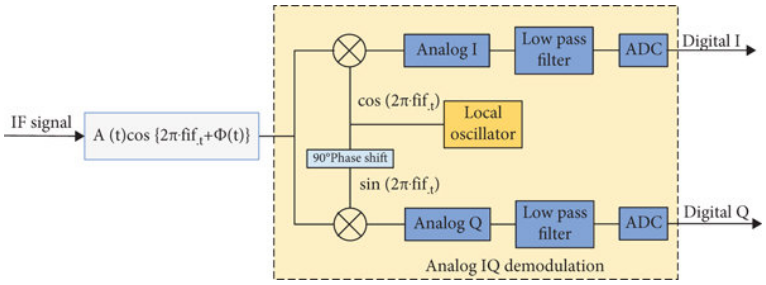


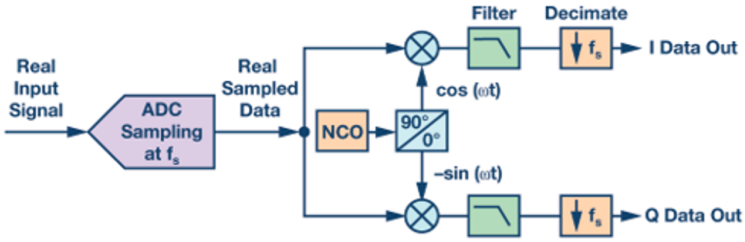
Figure 2.5: An example for STC curve.

2.2.1.2 Downconversion

The signal received by radar is usually high frequency. Since it is difficult to process this signal directly in digital form, it needs to be reduced to a lower frequency. Baseband



(a) Analog IQ demodulation



(b) Digital IQ demodulation

Figure 2.6: IQ demodulation.

downconversion allows the signal modulated at the carrier frequency to be separated from the carrier and represented at a lower frequency.

A signal can be downsampled to baseband in analog or digital. If the signal is downsampled to baseband in analog, two mixers and two ADCs are required. However, if the signal is sampled at IF frequency and downsampled to digital baseband, one ADC is used. This baseband downconversion process is called IQ demodulation. Block diagrams of analog and digital IQ demodulation are shown in figure 2.16 [34].

The signal downconverted to baseband consists of two components, I and Q. I (In-phase) is in phase with the carrier signal, while Q (Quadrature) is in phase perpendicular to it. Subtracting these two signals allows a complex representation of the signal as in equation 2.2:

$$s(t) = I(t) + jQ(t) \tag{2.2}$$

This preserves the amplitude and phase information in the signal. In this manner, range, Doppler and direction information of the target can be calculated using the amplitude and phase information of the signal.

2.2.1.3 Filtering

The signal downconverted to the baseband needs to be filtered at the frequency of interest. One of the most commonly used filtering methods for this purpose is Finite Impulse Response (FIR) filters. FIR filters are applied in the baseband to pass certain frequency components of the radar signal and suppress unwanted frequency components. This also suppresses the noise in the signal. At this stage, decimation can also be applied to reduce the data rate.

FIR filters are generally preferred in radar applications because their linear phase properties ensure that the time and phase information of the signal is not distorted. Moreover, they are structurally stable, relatively easy to design and compatible with digital processing hardware, which increases their use in radar systems.

Furthermore, matched filtering is often the preferred technique to increase the target SNR. Its purpose is to optimize the correlation between the radar pulse that is transmitted and the received signal. This process provides the maximum SNR at the output. The matching filter is the time reversed and complex conjugate of the transmitted signal. If x is the transmitted signal, the matched filter is as seen in equation 2.3.

$$h(t) = \sigma x^*(T_M - t) \tag{2.3}$$

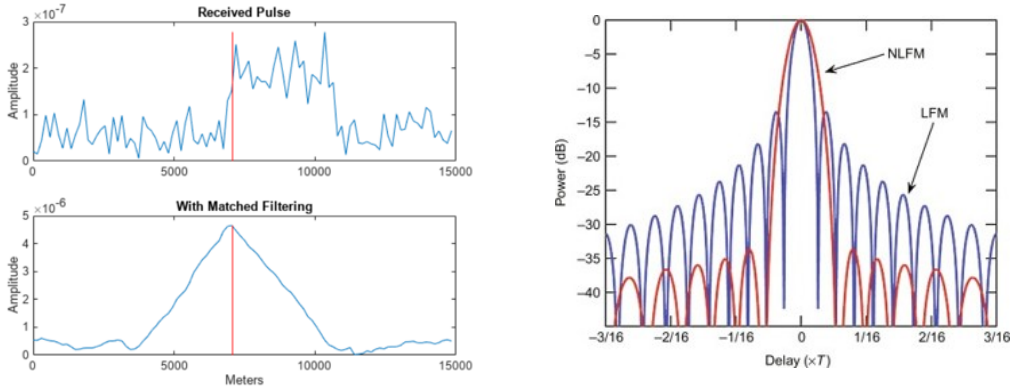
T_M describes the time at which SNR is maximized. If $x'(t)$ is the input signal that consists of target and noise, the output of the matched filter is determined as equation 2.4.

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} x'(\tau) h(t-\tau)\, d\tau \\ &= \int_{-\infty}^{\infty} x'(\tau) x^*(\tau + T_M - t)\, d\tau \end{aligned} \tag{2.4}$$

In this manner, even weak echoes from the target can be successfully detected in a noise.

Another method to increase SNR in radars is the pulse compression method. This method is based on the matched filter application of a phase or frequency coded pulse. Several different methods can be used to increase the SNR in an uncoded pulse, such as increasing the output power, increasing the pulse duration, and increasing the number of pulses. In order to increase the SNR in an uncoded pulse, several different methods can be used, such as increasing the output power, increasing the pulse duration, and increasing the number of pulses. However, this may not be possible due to hardware limitations and processing load. In addition, since range resolution is a parameter inversely proportional to the pulse duration, increasing the pulse duration will worsen the range resolution and this cause to have worse target separation.

However, in order to design long-range radars, it is also necessary to increase the pulse duration. Maximum gain is achieved by convolution of the received pulse with the conjugate and time inverse of the pulse code. Pulse code can be pulses such as Non-Linear Frequency Modulation (NLFM), Linear Frequency Modulation (LFM) where the frequency changes depending on time. In such pulse codes, the pulse is divided into N pulses. The pulse compression method increases the target SNR and improves the range resolution because it uses a wider bandwidth than the uncoded pulse. Figure 2.7 shows the matched filter outputs of the uncoded and coded pulses [35,36].



(a) an uncoded pulse

(b) a coded pulse

Figure 2.7: The matched filter outputs.

The matched filter output of the square wave is shown in figure 2.7 (a). As can be seen from the figure, the level of the side lobes is very high and there is a large spread in the target range. In figure 2.7 (b), the output of the matched filter belonging to the coded

signals is seen. One mainlobe and lower level sidelobes belonging to this mainlobe are seen. When the significant differences between NLFM and LFM are examined; it is seen that the sidelobe level of the NLFM signal in time is lower, but the mainlobe width is greater than LFM. The mainlobe width should be narrower for range resolution. In addition to the narrow mainlobe, the sidelobe level should also be low for detection performance. Although there is a loss in range resolution when NLFM is used instead of LFM, masking of low SNR targets can be prevented by the lower sidelobe level.

2.2.2 Slow-time processing

The operations performed between pulses are called slow time signal processing. These operations include noise suppression, MTI filtering and Doppler filtering. While Doppler filtering is mostly performed to determine the target speed, it can also be used to suppress clutter in some applications. In applications where the target speed needs to be determined, Doppler resolution becomes an important parameter in order to correctly resolve the speed. Doppler resolution is determined by Pulse repetition frequency (PRF). PRF is inversely proportional to the pulse duration, i.e. pulse repetition interval (PRI). While a wider Doppler frequency range can be obtained by using a high PRF, a better Doppler resolution can be obtained by increasing the number of pulses.

Since pulses are integrated coherently in inter-pulse processing, it also increases the target SNR. This process is applied in the time-frequency domain. If N pulses are integrated coherently, the coherent integration gain becomes $10 * log(N)$ dB.

Clutter can be defined as unwanted signals reflected from fixed objects. One of the most basic methods to suppress clutter is MTI filtering. MTI filter suppresses the signals around DC and ensures that the data is purified from the fixed clutter in 0 Doppler. It can be shown MTI filter application in figure 2.8 [37].

Another way to separate the signal in the Doppler domain is Doppler filtering. For this, while the Fourier Transform (FFT) can be performed along the pulse axis of the signal, it is also preferable to use Doppler filter banks. Doppler filter banks are generally preferred in radar systems operating with low pulses, while FFT application is performed in systems without pulse restrictions. An example Doppler filter bank is seen

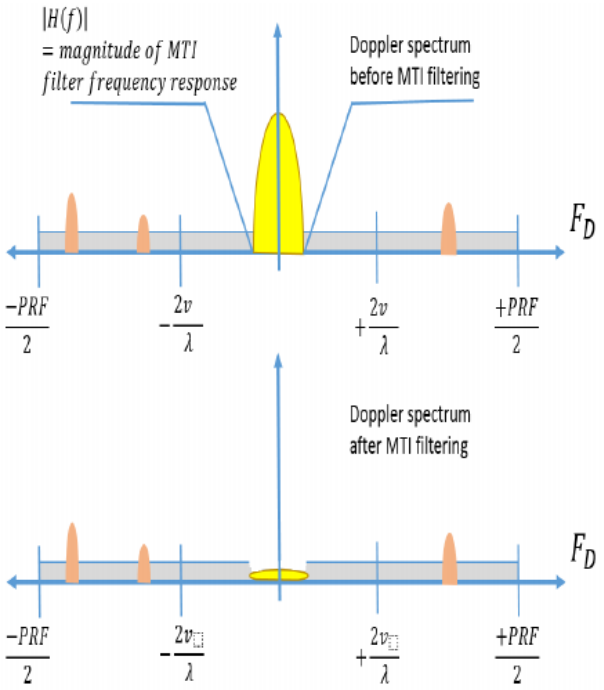


Figure 2.8: The MTI filtering.

in Figure 2.9 [38]. The general purpose of these filters is not to obtain target velocity information, since the Doppler resolution will be low when low pulses are used, but to suppress the clutter.

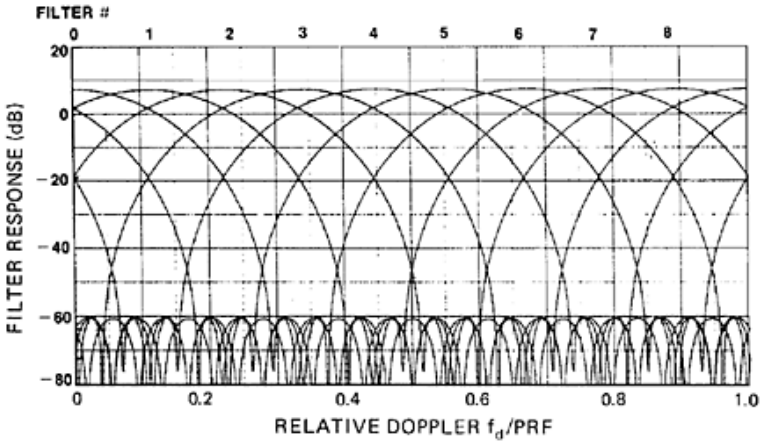


Figure 2.9: Doppler filters bank.

2.3 Radar Detection Algorithms

2.3.1 Traditional approaches

Essentially, radar target detection is treated as a binary hypothesis testing problem. With this method, the radar system tries to distinguish between whether the received signal contains only noise (H0: no target) and whether it contains both target signal and noise (H1: target present). This detection decision is usually made by comparing the signal at the receiver output with a certain threshold level. If the receiver output is above the predetermined threshold level, it means there is a detection. Probability of detection (P_d) and false alarm probability (P_{fa}) are the two fundamental concepts that determines the system's performance. P_{fa} is the probability that the radar will mistakenly detect a target when none is there, whereas P_d is the probability that the radar will detect a target that is truly present. There is often a balance between these two concepts: P_d often increases P_{fa} as well. Therefore, radar systems aim to keep P_d as high as possible while keeping P_{fa} as low as possible. There is also a possibility that the signal may be evaluated as noise even though it contains a target. This situation is called missed detection. Although a high threshold reduces the probability of false alarms, it often increases missed detections. Statistical modeling of noise and clutter and threshold determination are used to achieve this balance [1]. Given that N data samples are used for detection and are grouped in the vector $y = [y_0 \ ... \ y_{N-1}]$ the PDFs would be represented as in equation 2.5.

$$\begin{aligned} p_y(y, H_0) &\text{ -> PDF of y given that target is absent} \\ p_y(y, H_1) &\text{ -> PDF of y given that target is present} \end{aligned} \tag{2.5}$$

Achieving this balance is not possible in the case of setting a fixed threshold due to changing environmental conditions. Therefore, an adaptive threshold must be set depending on the environment. CFAR algorithms are traditional and widely employed techniques in radar target detection, designed to maintain a constant false alarm rate despite varying noise and clutter conditions. This adaptability is vital for achieving reliable target detection, especially in dynamic and complex radar environments [39].

CFAR methods operate by estimating the local noise level around the cell under test (CUT) using neighboring reference cells and adaptively setting the detection threshold accordingly [40]. The cells immediately next to the cell under test are called guard cells. Guard cells are selected by considering the amount of spread of the target in order to prevent the threshold value to be calculated from being calculated incorrectly. In this way, the side lobes of the target are prevented from affecting the threshold value. Reference window cells are used to determine the noise level in the environment and a threshold value is calculated with the statistical data obtained from these cells and target detection is performed. Figure 2.10 shows the location of the cells belonging to the CFAR method.

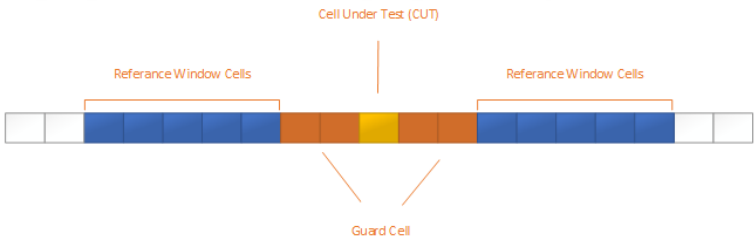


Figure 2.10: Location of cells for CFAR.

Among the classic CFAR variants are Cell-Averaging CFAR (CA-CFAR), Smallest Of Cell-Averaging CFAR (SO-CFAR), Ordered Statistics CFAR (OS-CFAR), and Greatest Of Cell-Averaging CFAR (GOCA-CFAR). CA-CFAR calculates the threshold by averaging the noise power of reference cells on both sides of the CUT, which works well in homogeneous noise but can be sensitive to interfering targets or clutter edges [41].

if N is considered as the number of cells in the reference windows and x_i is the data in the reference window cells, the estimation of the noise level E is calculated as in equation 2.6

$$E = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.6}$$

The estimation of noise E is multiplied by the scale factor K yields the detection threshold T. The general expression of detection threshold is seen in equation 2.7.

$$T = EK \tag{2.7}$$

K parameter is calculated for desired P_{fa} value using the number of reference cells as in equation 2.8.

$$K = N\left(P_{FA}^{-1/N} - 1\right) \tag{2.8}$$

SOCA-CFAR improves robustness in the presence of clutter edges or multiple targets by selecting the minimum average noise power between the leading and lagging windows to set the threshold, thus reducing the risk of masking weaker targets. K is calculated for desired P_{fa} value as in equation 2.9.

$$P_{FA} = 2\sum_{k=0}^{N/2-1}\binom{N/2-1+k}{k}\left(2+\frac{K}{N/2}\right)^{-N/2-k} \tag{2.9}$$

The detection threshold is expressed is as in equation 2.10

$$T = Kmin(E_{lagging}, E_{leading}) \tag{2.10}$$

Conversely, GOCA-CFAR selects the maximum average noise power from these windows, which is useful in scenarios where clutter is dominant on one side. K is calculated for desired P_{FA} value as in equation 2.11.

$$P_{FA} = 2\left(1+\frac{K}{N/2}\right)^{-N/2} - 2\sum_{k=0}^{N/2-1}\binom{N/2-1+k}{k}\left(2+\frac{K}{/2}\right)^{-N/2-k} \tag{2.11}$$

The detection threshold is expressed is as in equation 2.12

$$T = Kmax(E_{lagging}, E_{leading}) \tag{2.12}$$

OS-CFAR enhances performance in non-homogeneous environments by sorting the reference cells' noise levels and choosing a ranked statistic (often the minimum or a specific order statistic) as the threshold. This approach is particularly effective in mitigating false alarms caused by localized interference or clutter [42]. K is calculated for desired P_{FA} value as in equation 2.13.

$$P_{FA} = k\binom{N}{k}B(K+N-k+1,k) \tag{2.13}$$

where $B(.)$ is the beta function and k is the rank. The detection threshold T is calculated by multiplication of K and calculated E above.

Due to their adaptability and computational simplicity, CFAR methods remain fundamental in radar signal processing and often serve as baseline benchmarks when evaluating emerging detection techniques such as machine learning and deep learning models.

2.3.2 Deep learning based approaches

Traditional approaches often struggle in complex environments characterized by non-stationary clutter, low SNR, and heterogeneous interference. Deep learning techniques have recently gained significant attention for their ability to automatically learn hierarchical and discriminative features directly from raw or processed radar data [43]. Architectures such as CNN and Transformers can model intricate spatial and temporal dependencies, improving target detection performance beyond traditional methods.

2.3.2.1 Convolutional neural network

CNNs are specialized artificial neural network architectures that process image data in layers to classify or analyze it for regression purposes [44]. There are five main parts that explain the fundamental functions of the CNN structure such as an input layer, convolution layers, pooling layers, a fully connected layer, and an output layer [45]. The input layer is the first data source in the network's learning process, and it carries the image's pixel information. The subsequent convolution layer produces output by calculating the weights of the neurons connected to local regions in the input volume and the scalar product between these regions. In this layer, the basic features in the image (edge, texture, shape, etc.) are detected thanks to the learned filters. The ReLU (Rectified Linear Unit) activation function following the convolution layer deepens the network by applying a nonlinear transformation to the output of each neuron and provides the model with learning capacity. The pooling layer in the next stage performs a downsampling process along the spatial dimensions of the input. This process reduces the computational cost by reducing the size of the feature maps and provides

a more generalizable structure by limiting the overfitting tendency of the model. A fully-connected layers work as in standard artificial neural networks, converting the activations from the previous layers into class scores. The network matrix's output has been flattened and is prepared for the fully connected layer's classification procedure. These layers interpret the learned features and perform the classification task. In order to increase the performance, the ReLU activation function can be used between these layers. Thanks to this multi-layer transformation process, CNNs can produce effective classification and regression results by transforming the original image input layer by layer with convolutional and sampling-based operations [44]. A simple CNN architecture is shown in figure 2.11

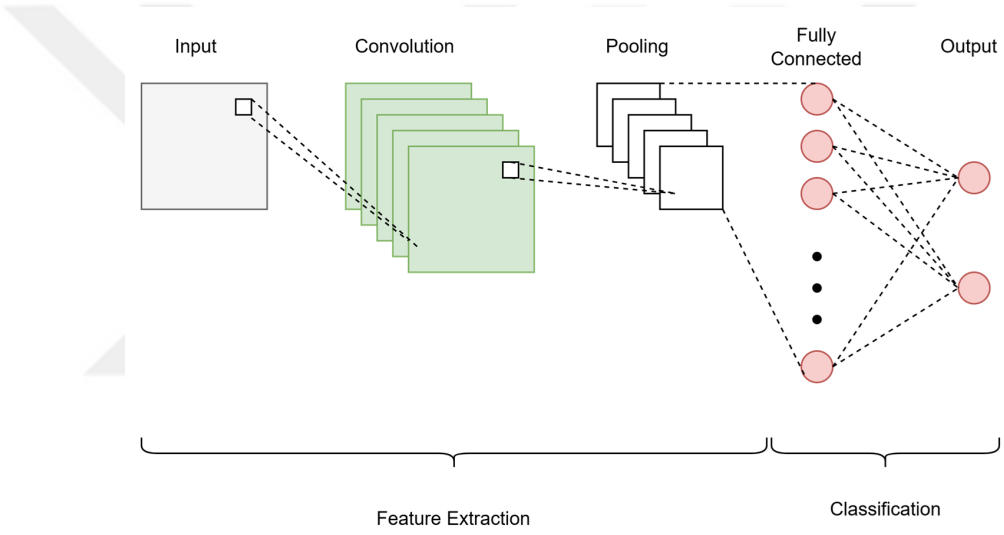


Figure 2.11: A simple CNN architecture.

In this study, CNN undertakes the task of data classification and feature extraction. CNN improves detection accuracy by processing the amplitude, shape, change points and other important information of the data. The general stages of the deep learning based detection process are shown in detail in Figure 2.12.

Given the range-Doppler matrix as input, the matrix is divided into smaller patches in the pre-signal processing step and the input to the CNN module is generated. Data feature extraction is performed by convolution layers in the CNN module. Then, with the pooling layer, the inferences are summarized by preserving the important information and a smaller sized data is obtained. Finally, the fully connected layer combines these

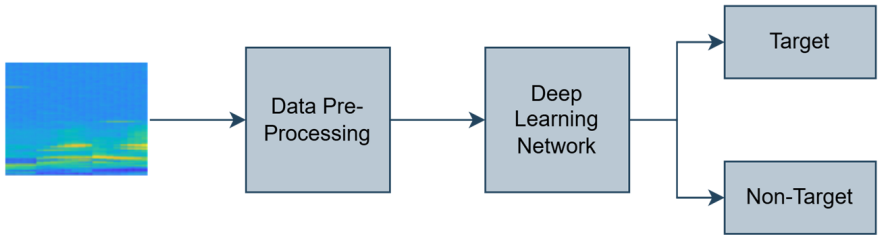


Figure 2.12: Deep learning detection structure.

features and performs classification. At the output of the CNN module, a target or non-target decision is made for the relevant patch. In order to compare the results, the model in [17] is taken as a reference, where two convolution layers are used and a pooling layer is added between the convolution layers. These layers are followed by three fully connected layers with a dropout layer in between. The model used is visualized in detail in Figure 2.13.

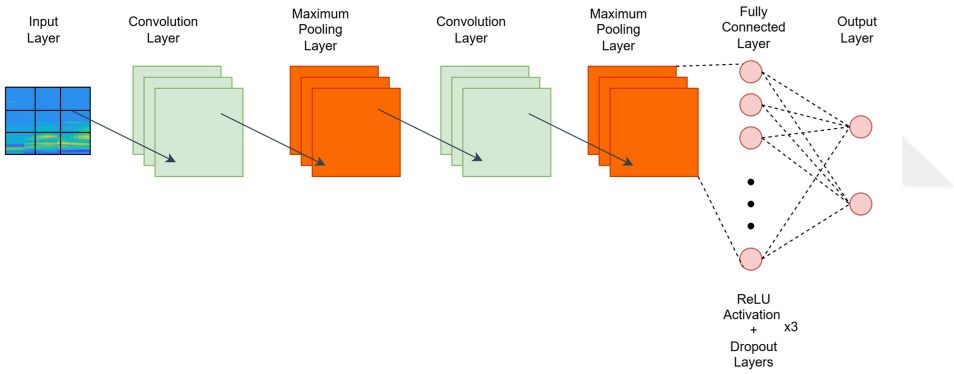


Figure 2.13: CNN architecture.

CNNs are often used to learn spatial structures and extract local features from images. However, because CNNs typically focus on pixels in a small area, they struggle to capture connections between larger and more distant regions in the image. As a result, the understanding of global context is limited. To overcome this problem, transformer-based models such as ViT break the image into small patches, analyze them one after the other, and use attention mechanism to learn long-range relationships between these patches. As a result, transformer-based methods are better able to convey the complex connections between distant areas of the image.

2.3.2.2 Vision transformer

Transformers are deep learning based architectures widely used in NLP. ViT is their adaptation to visual data. CNNs are unable to accurately model long-range relationships over the entire image due to their limited sensing range. ViT tries to solve this limitation by using the self-attention mechanism that can simultaneously model the interactions between components in the image. Unlike classical neural networks, ViT processes images not directly at the pixel level, but by decomposing them into small fixed-size patches [46].

In figure 2.14, it can be shown standard ViT structure and In figure 2.15, it can be shown multi-head attention mechanism.

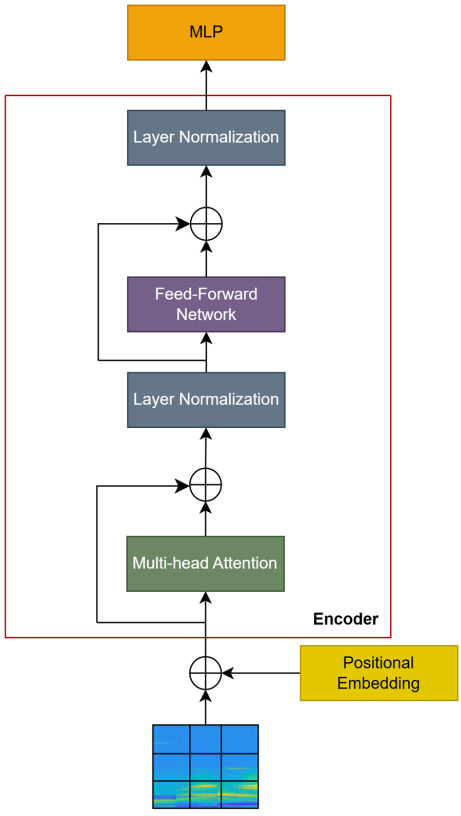


Figure 2.14: ViT structure.

The classical transformer structure consists of two main parts: Encoder analyzes the input data and creates meaningful representations, while Decoder processes this

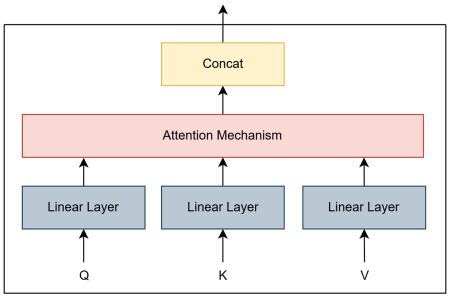


Figure 2.15: Multi-head attention mechanism.

information and produces the final output. However, the basic structure of ViT consists of blocks of encoders, which are used to learn the contents of the patches and their location in the image. ViT only knows which part contains what, but not where it belongs in the image. This makes it difficult to understand visual patterns such as shapes, edges and structures. Therefore, positional embeddings are a mandatory component for spatial awareness in patch-based structures such as ViT. To preserve positional information, position embeddings are added. The encoder takes the resulting embedding vector sequence as input [25]. Multi-head attention mechanism and Feed-Forward Neural Network (FFN) or MLP, are the two main components in each encoder block. Each patch interacts with all other patches via the Multi-head attention mechanism to distribute attention. This multi-headed structure can model different types of relationships in parallel through various headers. By applying non-linear transformations to each patch representation, the subsequent MLP layers increase the learning capacity of the model.

The self-attention mechanism provides an effective transformation by learning contextual relationships through the query (Q), key (K) and value (V) matrices obtained from the input data. The attention function is given in equation 2.14. If Q, K and V are derived from the same input data, this process is called self-attention.

$$Attention(Q,K,V) = softmax(QK^T/\sqrt{d_k})V \tag{2.14}$$

Multi-head attention mechanism processes information from different perspectives in parallel and combines the obtained results. In this way, the model can learn different aspects of the data simultaneously. This is implemented as shown in equation 2.15.

$$MultiHead(Q,K,V) = Concat(head_1,...head_h)W^O$$
$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.15}$$

The matrices W_i^Q,W_i^K,W_i^V in the equation represent the query, key and value transformations, respectively. Their dimensions are given as d_mxd_q, d_mxd_k, d_mxd_v, respectively. W_i^O, which performs the output mapping, has the dimension hd_vxd_m. While h denotes the number of heads for the multi-head attention mechanism, d_m denotes the size of the encoding vector [21].

Layer Normalization and residual connections are located between and at the end of these Multi-head attention mechanism and MLP components. In this way, the model uses both the original knowledge and the new learned representation at each step. These components facilitate the reduction of information loss, more stable learning and deepening of the network.

The attention-based structure of ViT is more flexible in learning global context relations than CNNs. Thanks to this feature, ViT exhibits strong performance, especially on large datasets and with sufficient pre-training. ViT derivatives have been successfully adapted to more complex computer vision tasks such as object detection, segmentation and medical image analysis as well as image classification.

2.3.2.3 Swin transformer

Feature extraction in ViT produces feature maps at a single resolution. Associated with the size of the input image, self-attention has a high computational complexity as it is computed globally across the entire image. Swin Transformer is an architecture designed to overcome the computational burden and local context capture constraints of ViT. In Swin Transformer, self-attention is performed in shifted and fixed-size windows. This makes self-attention calculations much simpler and allows the model to work well with high-resolution visuals.

Each of the small, fixed-size patches of the input image is transformed by a linear layer into a low-size embedding. The self-attention is then computed within predefined, fixed-size windows generated by these embeddings. However, this window-based approach can limit the interaction between window boundaries. Windowing may

cause information loss during window transitions. To solve this problem, in the Swin Transformer architecture windows are shifted by half a window in each layer. This shifting mechanism allows different window areas to cross-interact with each other, enabling better modeling of both local and international context information.

Swin Transformer is constructed by replacing the standard multi-head self-attention module (MSA) in a Transformer block with a module based on shifted windows. It's kept the other layers the same. As shown in Figure 2.16, a Swin Transformer block consists of a window-based MSA module (W-MSA) and shifted window-based MSA module (SW-MSA) followed by a 2-layer MLP between them respectively. A Layer Normalization (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module.

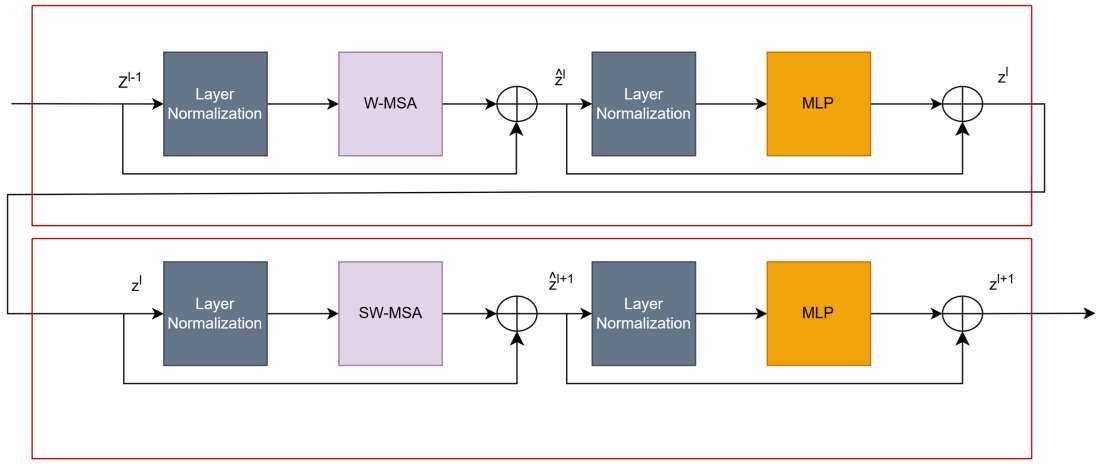


Figure 2.16: Swin transformer block.

Since there are no connections between windows, the modeling capabilities of the window-based self-attention module are limited. In successive Swin Transformer blocks, a shifted window partitioning technique that alternates between two partitioning configurations is suggested. The equation 2.16 is used to determine consecutive Swin Transformer blocks using the shifting window segmentation approach.

$$
\begin{aligned}
\hat{z}^{l} &= \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \\
z^{l} &= \text{MLP}(\text{LN}(\hat{z}^{l})) + \hat{z}^{l} \\
\hat{z}^{l+1} &= \text{SW-MSA}(\text{LN}(z^{l})) + z^{l} \\
z^{l+1} &= \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1}
\end{aligned}
\tag{2.16}
$$

The output features of the W-MSA and SW-MSA module and the MLP module for block l are indicated by the symbols $hatz^{l}$ and z^{l} respectively. Window-based multi-head self-attention is indicated by W-MSA and SW-MSA, which use direct window partitioning and shifted window partitioning modules respectively [26].

The shifted window segmentation method creates links between adjacent non-overlapping windows in the preceding layer and has shown effective in picture classification, object detection, and semantic segmentation.

3. PROPOSED METHODS

3.1 Improved Vision Transformer

In this study, a new target detection method is proposed to overcome various limitations observed in existing approaches such as the high sensitivity of traditional methods to environmental conditions, the inability of CNN-based models to effectively extract global contextual features, the need for large datasets for ViT and Swin Transformer models, and the relatively weaker ability of Swin Transformer to extract and generalize global features compared to the ViT model.

Transformer model is a deep learning method that learns with self-attention mechanism by giving variable weights to different components of the data. In this way, it evaluates the importance of each input in the context and processes the information more effectively by capturing long-term relationships. In the proposed method, instead of the Multi-Layer Perceptron (MLP) layer in the classical ViT structure, an improved feedforward network structure including a convolution layer is used. In this way, it is aimed to detect targets more accurately in the presence of chaos.The structure of the proposed method is shown in Figure 3.1.

The first steps of the structure are similar to the classical transformer. Position embedding is done on the range-Doppler matrices taken as input so that the transformer model can understand the sequential data. With position coding, a position information is added to each data and thus the model is enabled to learn the spatial relationships between the data.

Since ViT has a more complex structure than CNN architecture, it needs to work with larger datasets. However, due to insufficient data resources, this thesis proposes new radar target detection methods based on the improved ViT architecture to achieve higher performance with less data. In the proposed method, especially in complex radar environments, which can limit the generalization capacity of the model, instead of the

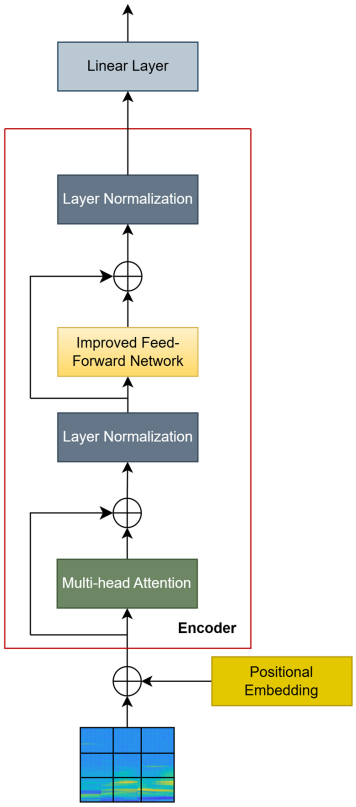


Figure 3.1: Improved ViT architecture.

Multi-Layer Perceptron layer in the classical transformer model, an improved advanced network structure is used [47].

Detailed explanation of improved feed-forward network is shown in figure 3.2.

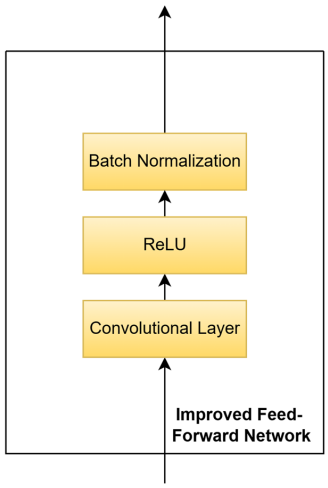


Figure 3.2: Improved feed-forward network.

In this way, the features in the image data are better captured and the difference between the clutter and the real targets can be determined more clearly. This layer used includes a convolution layer and ReLU (Rectifier Linear Unit) is preferred as the activation function. Then, batch normalization is applied and the activations in the convolution layers are normalized, thus ensuring that the model learns more stably and quickly.

3.2 Hybrid Model

The hybrid model used in this work aims to combine the advantages of CNN and Transformer-based architectures. Especially in SAR images, the combination of convolutional and Transformer structures has become increasingly common in order to capture local details accurately and to model global spatial contexts. In this context, the main inspiration for the hybrid model structure is the CTF-Net (Convolutional and Transformer Fusion Network) architecture proposed by Wu et al. [48]. The hybrid model applied for SAR images in the aforementioned study will be applied for raw radar data in this thesis. The proposed detection model is illustrated in figure 3.3.

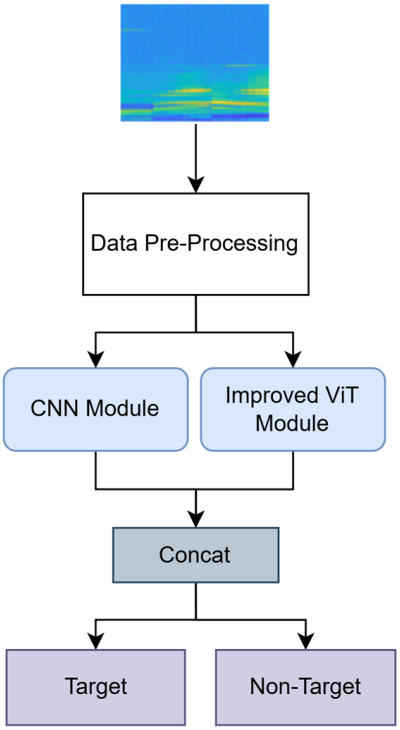


Figure 3.3: The hybrid structure.

CTF-Net proposes a multi-layered fusion structure that combines CNN and Transformer structures in a scalable manner. The model proposed in the paper consists of CNN-Transformer pairs operating at three different scales and gradually fuses the outputs of these structures. At each scale, feature maps are processed in parallel in both convolutional and Transformer layers, and then fused at the end of each layer to integrate both local and global representation strengths. This multi-scale structure results in more effective feature representations for both small and large ship target detection.

In this thesis, a simplified version of the structure proposed by CTF-Net is implemented. Instead of the multi-layered and scaled CNN + ViT modules in the original model, only a single CNN and a single ViT block are configured in parallel without any scaling or inter-layer iteration due to the small data size. This structure is chosen to reduce complexity and lower computational cost.

Let the input image be represented as follows: $X \in \mathbb{R}^{H \times W \times C}$ where H and W are the height and width of the image and C is the number of channels. The convolutional branch extracts local features using standard convolutional layers. The output feature map can be defined as follows:

$$F_{\text{CNN}} = f_{\text{CNN}}(X) = \sigma(W_{\text{conv}} * X + b) \tag{3.1}$$

Where $*$ represents the convolution process, W_{conv} is the set of convolution filters, b is the bias term added to each feature map, and σ is a nonlinear activation function such as ReLU that adds a nonlinearity to the model and helps it to learn complex patterns. The convolutional branch is particularly effective in capturing local spatial features such as edges, textures and object boundaries by applying multiple filters across the image.

The CNN model used in the hybrid model is introduced in section 2.3.2.1.

On the other hand, the Vision Transformer (ViT) transforms the input image into flattened and linearly embedded vectors. These embedded patch tokens are further augmented with spatial encodings to preserve spatial information and then passed

through a stack of transformer encoder layers that utilize multi-head self-attention mechanisms to model global dependencies across the entire image.

$$F_{\text{ViT}} = f_{\text{Transformer}}(X) \tag{3.2}$$

The transformer structure used in the hybrid structure is the improved ViT structure mentioned in section 3.1.

In the hybrid model, the feature maps obtained from the input image are sent to both CNN and Transformer in parallel. Both structures process features in a way that is specific to their architecture. CNN strongly extracts local context and edge information, while Transformer provides a global view modeling longer-range correlations. At the end of the layer, the feature maps from both paths are concatenated to produce a rich representation that includes both local and global information as in equation 3.3;

$$F_{\text{fused}} = \text{Concat}(F_{\text{CNN}}, F_{\text{ViT}}, \text{dim} = C) \tag{3.3}$$

Since a single channel is used in the proposed model, C will be equal to 1. This combined feature map is passed on to subsequent classification or detection layers which consist of dense layers and ReLU activation layers.

This method improves model performance by utilizing the effect between different types of information representations for tasks such as target detection.

In conclusion, the hybrid model proposed in this thesis offers a leaner and more feasible structure while maintaining the basic principles of CTF-Net. Especially for applications with resource constraints or lower computational power requirements, this simplified hybrid model provides an effective alternative.

4. SIMULATION RESULTS

In this section, a comprehensive comparison is conducted between traditional target detection methods, specifically CFAR algorithms, and deep learning-based approaches, including CNN, ViT, Swin Transformer, improved Transformer architecture, and hybrid architectures that integrate CNN and improved ViT architectures. The evaluation emphasizes target detection performance, supported by simulations conducted on both synthetically generated and real-world datasets. Section 4.1 represents the simulation settings and parameters. Section 4.2 presents the results obtained from simulations using synthetic data and Section 4.3 presents the results derived from real-world data. Section 4.4 includes model complexities of deep learning methods.

4.1 Simulation Parameters

The range-Doppler radar data used in the simulations is represented as a two-dimensional matrix, where one axis corresponds to the range and the other axis corresponds to Doppler frequency which is relative velocity of the objects. This data, which serves as input to the detection algorithms, consists of 2400 range bins and 8 Doppler bins. For synthetic data simulations, it's created range-Doppler matrices that include twenty targets which is spread almost ten range bins. Synthetic data simulation parameters are demonstrated in table 4.1. An S-band pulsed radar is considered in the simulations. Accordingly, the frequency is selected in S-band and PRF and PRI is selected depending on the operational range. PRF is selected as 700 Hz and PRI is used as 1.4 ms accordingly. 8 pulses are used within a beam duration and a coherent integration gain of 8 pulses is added. A long pulse of 100 us duration is used and a short pulse of 1 us duration is used to cover the blind range caused by the long pulse. NLFM is used as the pulse code. The generated targets have speeds ranging between 0 and 400 knots and different locations between 0 and 60NM. Target SNR values are given different values between 20dB and 50dB.

Table 4.1: Synthetic data simulation parameters.

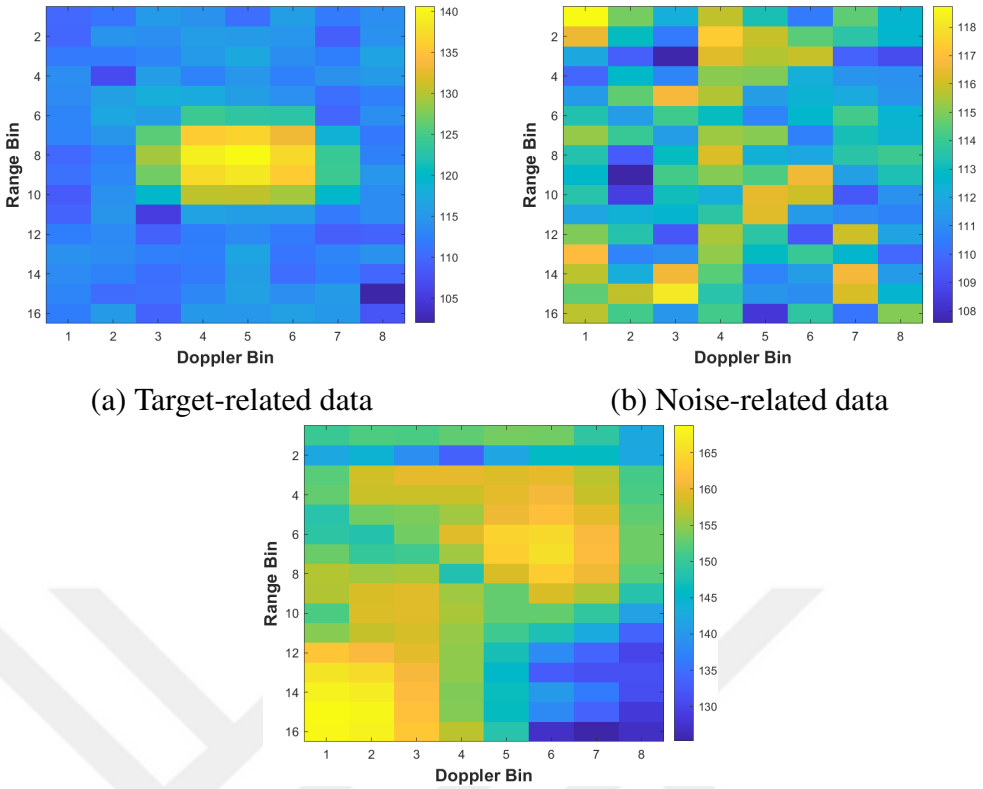
Parameter	Value
Frequency	S band
PRI	1.4 ms
PRF	700 Hz
Number of Pulses	8
τ_{short}	1 us
τ_{long}	100 us
SNR	20-50 dB
Velocity	0-400 knot
Pulse Code	NLFM
Range	0-60 NM

To prepare this data for deep learning applications, the range-Doppler data is divided into smaller, more manageable patches. Taking into account the spatial spread of a target along the range dimension, the data was segmented into patches of 16 range bins, while the entire Doppler axis was included in each patch. Patch extraction process was performed by sliding a window over the range axis with a stride of one range bin, preserving the sequential order of the data. Consequently, small range-Doppler matrices of size 16×8 were generated to serve as inputs for the detection algorithms. Since supervised learning methods require labeled data, each patch is annotated to indicate the presence (label 1) or absence (label 0) of a target. Figure 4.1 illustrates example patches corresponding to target and non-target data.

4.2 Detection Performance with Synthetic Data

The synthetic dataset is composed exclusively of noise and target signals, with targets modeled under the Swerling-0 assumption, indicating a constant radar cross-section (RCS) and no temporal fluctuation in the reflected signal. Data were generated to span a range of signal-to-noise ratios (SNRs) from 20 dB to 50 dB, incorporating various target velocities and ranges.

For each simulation scenario, a dataset containing 5,000 patches is used, randomly split into training and testing sets at an %80-%20 ratio respectively. A total of 4,000 samples were allocated for training, with 1,500 patches containing targets and 2,500 patches without targets. In simulations using synthetic data, the target-free patches consisted solely of noise, whereas in simulations with real data, the target-free set included both noise and clutter. The target patches represent aerial vehicles moving at varying ranges and angles relative to the radar, traveling at either high or low speeds.



(a) Target-related data

(b) Noise-related data

(c) Clutter-related data

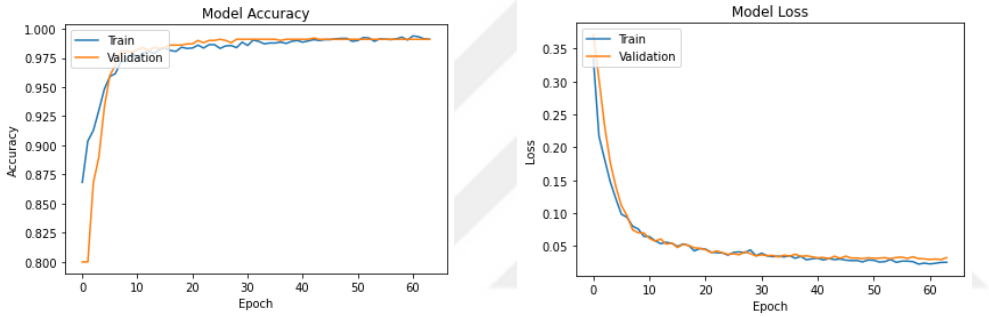
Figure 4.1: Range-Doppler patches corresponding to different data classes.

During the CFAR simulations, the probability of false alarm (P_{FA}) was fixed at 10^{-6}, and detection thresholds were accordingly set to maintain this false alarm rate. In the OS-CFAR algorithm, the variant selecting the minimum value between the leading and lagging reference windows adjacent to the target is designated as OS_{min}, whereas the variant choosing the maximum value is referred to as OS_{max}. The CNN architecture employs convolutional layers with kernel sizes of 32 and 64, respectively, followed by pooling layers with a kernel size of 2×2 for spatial downsampling. Each of the deep learning models used uses Adaptive Moment Estimation (ADAM) optimizer to update their weights during training.

Deep learning experiments were conducted using 64 training epochs with a 16 fixed batch size for all models. For CNN and Transformer-based architectures, different learning rates were used to reduce the risks of overfitting and underfitting and to achieve optimal convergence. Specifically, Transformer models were used with a reduced learning rate of 0.0001, while CNN models were used with a learning rate of 0.001.

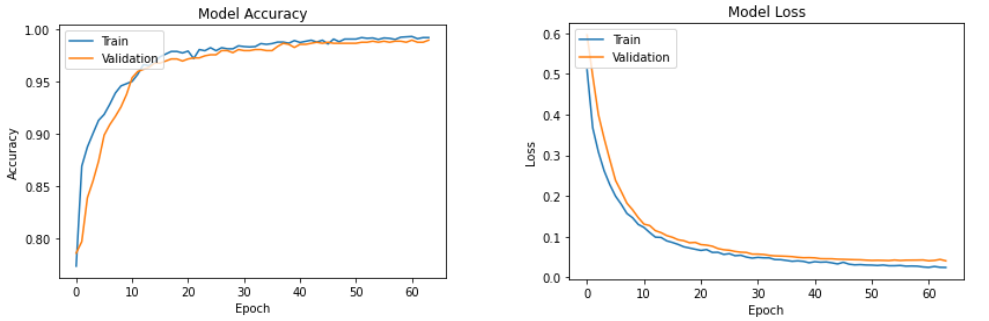
This choice was made based on experimental results collected during training. The CNN model demonstrated stable convergence at a learning rate of 0.001, as evidenced by consistent and simultaneous decreases in training and validation loss and convergence of accuracy values in both datasets. However, due to their deeper and more complex topologies, Transformer-based models exhibited greater sensitivity to the learning rate. These models tended to exhibit instability or suboptimal convergence behavior at higher learning rates.

When using a lower learning rate of 0.0001, Transformer models were able to achieve CNN-like convergence in experiments, prevent overfitting, and significantly reduce validation loss. This learning rate was selected for Transformer-based designs to ensure better generalization performance and consistent optimization dynamics.



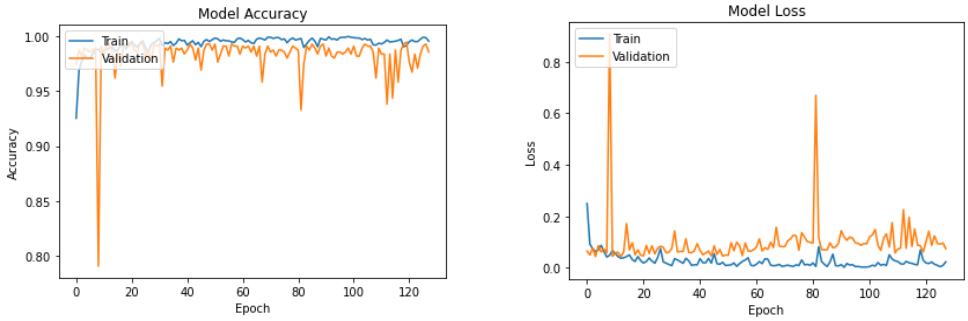
(a) Model training accuracy (b) Model training loss

Figure 4.2: CNN model accuracy and loss figures.



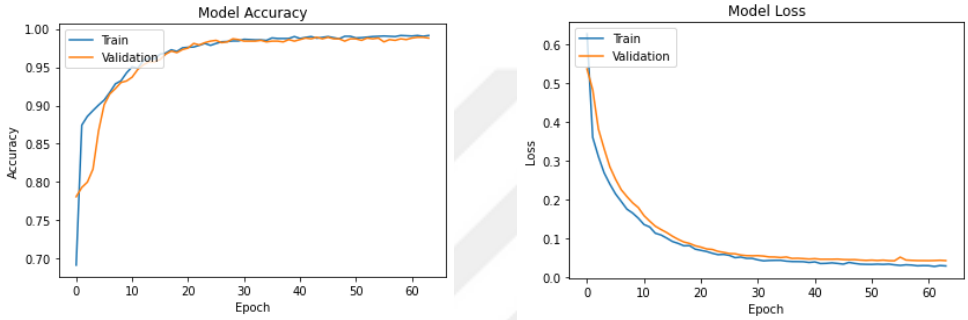
(a) Model training accuracy (b) Model training loss

Figure 4.3: ViT model accuracy and loss figures.



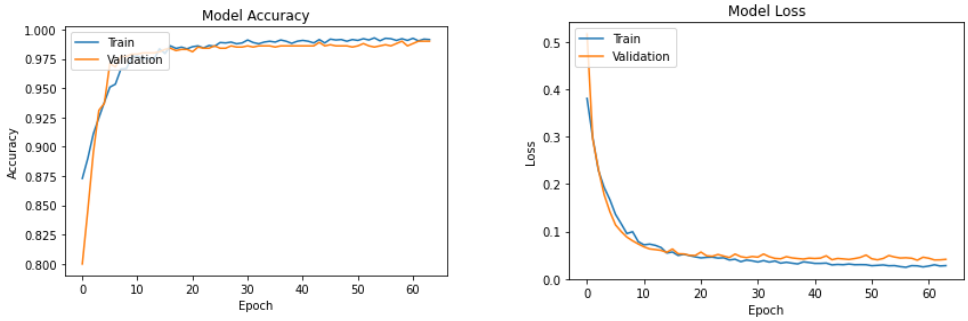
(a) Model training accuracy (b) Model training loss

Figure 4.4: Swin transformer model accuracy and loss figures.



(a) Model training accuracy (b) Model training loss

Figure 4.5: Improved ViT model accuracy and loss figures.



(a) Model training accuracy (b) Model training loss

Figure 4.6: Hybrid model accuracy and loss figures.

The training accuracy and loss curves of the models over the 64 epochs are shown in following figures. Figure 4.2 shows the training and validation accuracy and loss curves of the CNN model. It is clear that the loss curves decrease with minimal change, while the accuracy and loss curves converge and gradually increase. This curves show that the data provides generalizable features and that there is no overfitting or underfitting during

training. The convergence of these indicators validates the stability and effectiveness of the selected training configuration and learning rate.

The accuracy and loss curves of the basic Vision Transformer model are shown in Figure 4.3. Performance metrics for training and validation are progressing consistently, as in CNN. The convergence of accuracy and loss values indicates that the model has been successfully trained without overfitting or underfitting. Using a lower learning rate of 0.0001 helps stabilize the training dynamics of ViT, prevent oscillations, and promote a steady progression, as observed in the graphs.

Figure 4.4 shows the training and validation accuracy and loss curves of the Swin Transformer model for training on simulated radar data. It is seen that, compared to other models, it is prone to memorizing data and there is more oscillation between epochs. Due to the inadequacy of the data set or architecture of swin transformer model, it is unable to make sufficient generalizations and becomes more prone to overfitting. It is observed that accuracy and loss values show a sudden increase in some epochs.

In figure 4.5, the improved Vision Transformer model exhibits well-behaved training and validation trajectories. The loss curves decrease steadily and approach similar minimum values, while the accuracy curves show close agreement, indicating consistent generalization between training and validation data. The improvements applied to the ViT architecture appear to enhance convergence stability without causing overfitting or underfitting, thereby validating the effectiveness of the changes and hyperparameter settings employed.

Figure 4.6 shows the performance metrics of the hybrid model that combines convolutional and transformer-based components. Accuracy and loss curves are close to each other for both training and validation sets, indicating that the model benefits from the complementary strengths of each architecture. There are no indications of overfitting or underfitting, and the model achieves smooth convergence. These results confirm that the hybrid approach maintains strong generalization capabilities while preserving training stability.

The accuracy results for synthetic data corresponding to CA, SO-CA, GO-CA, OS_{min}, OS_{max}, CNN, ViT, Swin Transformer and the proposed methods are presented in table 4.2.

Table 4.2: Detection accuracy rates using simulated data.

Model	Accuracy(%)
CA	66.56
SO-CA	95.24
GO-CA	64.40
OSmin	86.60
Osmax	97.16
CNN	98.76
ViT	98.16
Swin Transformer	98.60
Improved ViT	98.80
Hybrid Model	**98.91**

Among the CFAR algorithms, the OS-CFAR method achieves superior detection performance. This improvement is primarily due to its robustness in scenarios involving closely spaced targets, where the threshold selected by OS-CFAR is less likely to overshadow or mask targets that have low SNR values a common issue observed with other CFAR variants. Specifically, OS-CFAR minimizes the missed detection of weak targets by adaptively choosing threshold values from reference windows adjacent to the target, thereby reducing interference effects. In the comparative evaluation between the CNN-based model, ViT, Swin Transformer and the proposed improved ViT and hybrid method, detection accuracies are found to be closely matched. However, the proposed methods consistently delivers marginally higher detection rates, indicating a slight but meaningful enhancement in target discrimination capability.

It can also be observed from the table that the classical ViT model exhibits lower performance compared to the CNN. ViT models have demonstrated success in capturing complex patterns in images. However, when radar data lacks such complexity or the available data is insufficient, ViT may struggle to effectively learn the underlying features. Consequently, in these scenarios, CNNs may achieve higher performance compared to ViT due to their ability to better generalize from limited or less complex data.

The categorical cross entropy loss function, which is commonly used in multi-class classification problems, was chosen for training the model. This function calculates how well the probability distribution predicted by the model matches the one-hot encoded

actual data labels. Thus, the probability of the correct class is maximized while the probabilities of the incorrect classes are minimized.

The categorical cross entropy loss function for a single sample is calculated as equation 4.1.

$$L = -\sum_{j=1}^{C} y_j \log(\hat{y}_j) \tag{4.1}$$

The C in the formula represents the total number of classes. y_j is the jth element in the one-hot encoding of the actual label of the data; this value is 1 for the correct class and 0 for all other classes. $\hat{y}_j$ represents the probability of the jth class predicted by the model using the softmax function. Thus, the loss function enables learning by calculating how high the probability is that the model decided on the correct class.

Loss values for simulated data have been calculated separately for training and testing. Table 4.3 shows the values of training loss.

Table 4.3: Training loss of deep learning models using synthetic data.

Model	Training Loss
CNN	0.01761
ViT	0.01825
Swin Transformer	0.01844
Improved ViT	0.01862
Hybrid Model	**0.01581**

Although the loss values for each model are similar, the lowest loss value was calculated for the hybrid model.

Table 4.4 shows the values of testing loss.

Table 4.4: Testing loss of deep learning models using synthetic data.

Model	Testing Loss
CNN	**0.04746**
ViT	0.06095
Swin Transformer	0.05839
Improved ViT	0.05285
Hybrid Model	0.04768

Although the loss values for each model are close to each other, it can be seen that the CNN model has the lowest loss value with a very small difference.

4.3 Detection Performance with Real-World Data

Real radar measurements encompass complex environmental factors such as clutter and noise, providing a more challenging and realistic test scenario. Results obtained from real radar data demonstrate the practical applicability and effectiveness of the proposed methods in operational settings.

For each simulation scenario, a dataset containing 5,000 patches is used, randomly split into training and testing sets at an %80-%20 ratio respectively. A total of 4,000 samples were allocated for training, with 1,500 patches containing targets and 2,500 patches without targets. In simulations using synthetic data, the target-free patches consisted solely of noise, whereas in simulations with real data, the target-free set included both noise and clutter. The target patches represent aerial vehicles moving at varying ranges and angles relative to the radar, traveling at either high or low speeds.

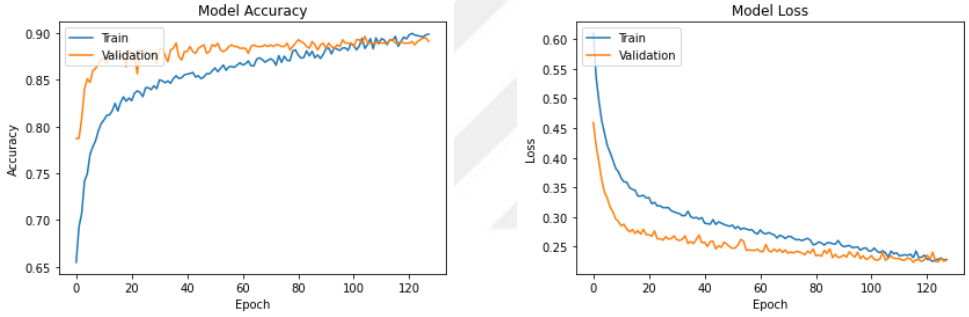
During the CFAR simulations, the probability of false alarm (P_{FA}) was fixed at 10^{-6}, and detection thresholds were accordingly set to maintain this false alarm rate. In the OS-CFAR algorithm, the variant selecting the minimum value between the leading and lagging reference windows adjacent to the target is designated as OS_{min}, whereas the variant choosing the maximum value is referred to as OS_{max}. The CNN architecture employs convolutional layers with kernel sizes of 32 and 64, respectively, followed by pooling layers with a kernel size of 2×2 for spatial downsampling. Each of the deep learning models used uses Adaptive Moment Estimation (ADAM) optimizer to update their weights during training.

Compared to simulations using synthetic data, simulations using real radar data required more training periods. This is due to the higher complexity and unpredictability seen in real-world observations. Under these conditions, the number of epochs was increased to 128 to ensure that the training and validation accuracy and loss values converged steadily. This extended training has been confirmed to prevent overfitting and improve the models' ability to generalize and capture complex patterns in the data.

Models trained using real radar data exhibit greater loss values and lower accuracy than models built on synthetic data, per the examination of training and validation processes.

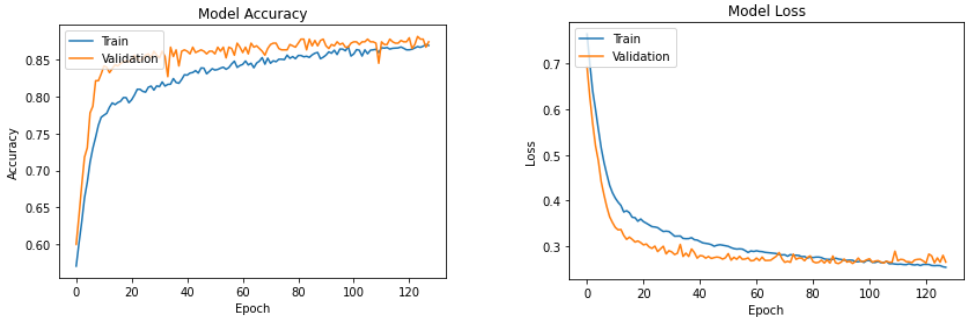
This discrepancy in performance is mostly caused by the complexity of real-world data, which has numerous sources of noise and clutter. The loss and accuracy curves show more fluctuations during training as a result. It can be said that makes learning process more challenging. The model's attempt to adjust to the varied and unexpected nature of real radar data is reflected in these oscillations.

Figure 4.7 illustrates the training and validation accuracy and loss curves of the CNN model for training on real radar data. Because of the greater variety and complexity of real-world signals, the model shows larger loss values and worse overall accuracy when compared to the synthetic data simulations. Despite this, there is no discernible divergence between the training and validation curves and there is no any signs to underfitting or overfitting. Given the nature of real data, some slight fluctuations are seen, which is to be expected.



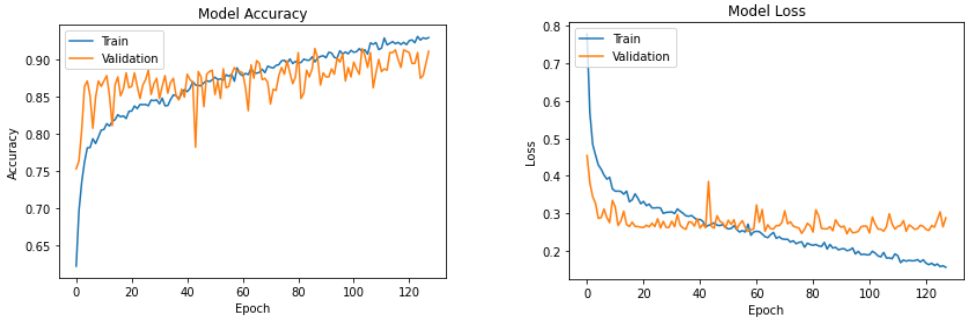
(a) Model training accuracy (b) Model training loss

Figure 4.7: CNN model accuracy and loss figures.



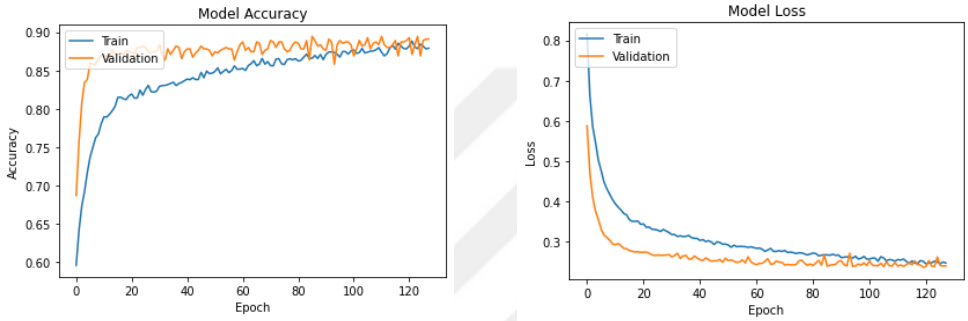
(a) Model training accuracy (b) Model training loss

Figure 4.8: ViT model accuracy and loss figures.



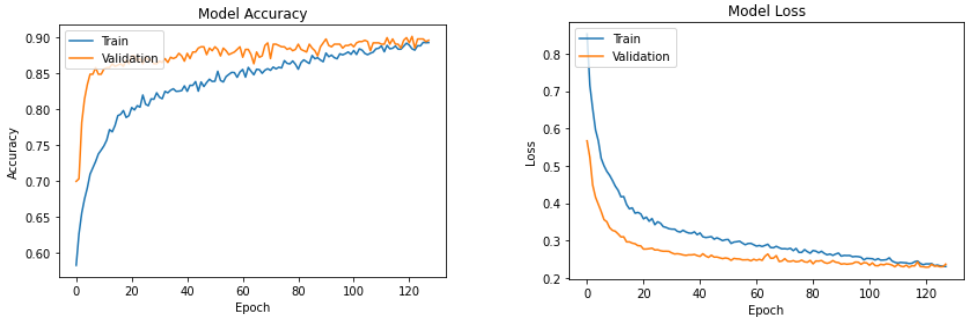
(a) Model training accuracy (b) Model training loss

Figure 4.9: Swin transformer model accuracy and loss figures.



(a) Model training accuracy (b) Model training loss

Figure 4.10: Improved ViT model accuracy and loss figures.



(a) Model training accuracy (b) Model training loss

Figure 4.11: Hybrid model accuracy and loss figures.

Figure 4.8 shows the training and validation accuracy and loss curves of the ViT model for training on real radar data. When compared with the CNN model, it is seen that the ViT structure converges more quickly. Similarly, overfitting and underfitting are not observed.

Figure 4.9 shows the training and validation accuracy and loss curves of the Swin Transformer model for training on real radar data. It is seen that, compared to other models, it is prone to memorizing data and there is more oscillation between epochs. It can be said that there is no stable learning process for this type of dataset.

Figure 4.10 shows the training and validation accuracy and loss curves of the improved ViT model for training on real radar data. The graphs are similar to the ViT model. Similarly, overfitting and underfitting are not observed.

Figure 4.11 shows the training and validation accuracy and loss curves of the hybrid model for training on real radar data. The convergence of the hybrid model took slightly longer than the convergence of the ViT and improved models. However, it can be said that learning process is completed successfully. Although it is a worse process than the simulated data, it can be evaluated that the model has acceptable accuracy and loss curves.

For the real data, the same number of training and testing samples as in the simulated data were used. While the simulated dataset contains only noise aside from the target data, the real dataset also includes clutter. The accuracy results for the real data corresponding to CA, SO-CA, GO-CA, OS_{min}, OS_{max}, CNN, ViT, Swin Transformer and the proposed methods are presented in table 4.8.

Table 4.5: Detection accuracy rates using real-world data.

Model	Accuracy(%)
CA	59.64
SO-CA	76.32
GO-CA	57.40
OSmin	72.48
Osmax	67.08
CNN	88.38
ViT	87.62
Swin Transformer	87.80
Improved ViT	89.42
Hybrid Model	**90.12**

Upon examining the results, it is evident that the presence of clutter in the dataset leads to a decrease in detection accuracy across all methods. However, traditional CFAR algorithms exhibit higher false alarm rates compared to deep learning-based approaches under these conditions. When comparing the CNN model, ViT and Swin Transformer with the proposed methods, the improved ViT achieves a modest performance gain of approximately 1% in accuracy over the CNN and gain of approximately 2% in

accuracy over the ViT and Swin Transformer, while the hybrid model demonstrates an approximate 2% improvement over CNN and an approximate 2% improvement over ViT and Swin Transformer methods.

Loss values for real-world data have been calculated separately for training and testing. Table 4.6 shows the values of training loss. Due to the complex structure of the real data, it can be seen that the loss values are higher than those of the simulated data. It is deduced from the table that the model with the lowest training loss among all models is the swin transformer model.

Table 4.6: Training loss of deep learning models using real-world data.

Model	Training Loss
CNN	0.2529
ViT	0.2878
Swin Transformer	**0.1238**
Improved ViT	0.2772
Hybrid Model	0.2122

Table 4.6 shows the values of training loss. Unlike training loss, swin transformer have slightly maximum testing loss value. As it's seen from the table the lowest loss value in testing loss belongs to the hybrid model.

Table 4.7: Testing loss of deep learning models using real-world data.

Model	Testing Loss
CNN	0.2671
ViT	0.2979
Swin Transformer	0.3697
Improved ViT	0.2755
Hybrid Model	**0.2496**

4.4 Model Complexity

The term FLOP (Floating Point Operations) describes how many floating point arithmetic operations a model can perform. It is an important statistic for evaluating computational complexity, especially in deep learning models. It's often expressed as "FLOP" (Floating Point Operations Per Second). FLOP values determine the model complexity and indicates how much processing power it requires for inference or training. The computational cost of the model can be estimated. Higher FLOP values generally suggest deeper and more complex models, while lower FLOP values indicate models that can run faster and more efficiently.

As it's shown from the table 4.8, due to its use of window blocks, the Swin Transformer has less complexity as seen in ViT. When the model complexity is examined, it is seen

Table 4.8: Complexity of the deep learning models.

Model	FLOPs(M)
CNN	**1.267**
ViT	2.686
Swin Transformer	2.231
Improved ViT	2.950
Hybrid Model	4.999

that the model with the least complexity is CNN; the model with the most complexity is the hybrid model that combines CNN and improved ViT models.

5. CONCLUSION

Traditional approaches are based on predetermined statistical models and are sensitive to environmental changes, making target detection a challenging problem. To overcome these limitations, deep learning-based approaches have been used as an alternative to traditional methods. Although CNN, one of the deep learning methods, is capable of learning and identifying features in images, it is inadequate in capturing information in remote connections. To overcome this problem, transformer-based models are preferred. Among transformer-based approaches, although the Swin transformer model appears to be more efficient in terms of complexity, it may cause the model to learn insufficiently and lead to imbalances during training in data where global relationships are important, such as radar data, because self-attention is applied through windowing. Furthermore, since the Swin Transformer is a model that performs effectively on large datasets, it may be prone to overfitting when the dataset is insufficient, as it may not generalize adequately. This situation can lead to imbalances during training. Therefore, even though it is less complex, it is not possible to achieve much higher target detection performance with less data using the Swin Transformer model. Instead, there is a need for models that can better learn global relationships by considering both the range and Doppler axes in radar data and that have higher generalization capabilities. In this context, the Improved ViT and Hybrid models offer superior results in terms of both learning stability and accuracy, providing alternative solutions to this need.

Accurate and reliable target detection is of great importance in critical radar applications such as defense, air traffic control and search and rescue. In such applications, it is necessary to keep the false alarm rate low, not to miss real targets and to detect targets with high accuracy. For this reason, improvement in target detection probability can play a critical role in such radar applications.

In this study, an improved ViT structure and CNN-improved ViT Hybrid structure based approaches are proposed for radar target detection. Since ViT architectures are more

complex than CNN architectures, they need to work with larger datasets. However, due to insufficient data resources, this thesis proposes new radar target detection methods based on the improved ViT architectures to achieve higher performance with less data. In the improved ViT model, an improved feedforward network structure including convolution layer is used instead of MLP layer used in classical ViT structure. In the second model, a hybrid structure is designed by integrating the improved ViT architecture with a CNN in parallel.

When the model complexity is examined, it is observed that the hybrid model is the most complex as expected, while the CNN is concluded to have the lowest complexity. The hybrid model combining CNN and improved ViT structures was observed to have the highest complexity. Despite of this, as a result of the simulations, it is observed that the proposed methods provide higher detection performance compared to traditional methods or previously proposed CNN based approach especially for realistic scenarios including clutter noise. When the success of CNN in extracting local features and details is combined with the ability of ViT to learn long-range relationships and extracting global features it's demonstrated the hybrid use of these two models provide better performance than others.

In future work, lightweight versions of the proposed model architectures can be developed that can operate at lower computational costs while maintaining higher accuracy and overall performance. This will increase the model's applicability, particularly in real-time applications or resource-constrained systems. The model can be optimized using architectural simplifications that reduce complexity, parameter reduction techniques, or information compression methods. This could enable the work to have a broader impact in both academic and practical fields.

REFERENCES

[1] **Richards, M.A.** (2014). *Fundamentals of Radar Signal Processing*, McGraw-Hill Education.

[2] **Burger, J.**, **Gochfeld, M.**, **Jeitner, C.**, **Burke, S.**, **Stamm, T.**, **Snigaroff, R.**, **... and Weston, J.** (2007). Mercury levels and potential risk from subsistence foods from the Aleutians, *Sci Total Environ*, *384*(1-3), 93–105.

[3] **Skolnik, M.I.** (1981). *Introduction to RADAR Systems*, McGraw-Hill Education.

[4] **Wang, M. and Chan, A.K.** (2005). Radar Signal Detection, *Encyclopedia of Electrical and Electronics Engineering*, `https://doi.org/10.1002/0471654507.eme345`.

[5] **Levanon, N. and Mozeson, E.** (2004). *Radar Signals*, Wiley-Interscience, Hoboken, NJ.

[6] **Lin, C.H.**, **Lin, Y.C.**, **Bai, Y.**, **Chung, W.H.**, **Lee, T.S. and Huttunen, H.** (2019). DL-CFAR: A Novel CFAR Target Detection Method Based on Deep Learning, *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA, pp.1–6.

[7] **Yavuz, F.** (2021). Radar Target Detection with CNN, *Proceedings of the 2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, pp.1581–1585.

[8] **Krizhevsky, A.**, **Sutskever, I. and Hinton, G.E.** (2012). ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., pp.1097–1105.

[9] **Kłosowski, P.** (2018). Deep Learning for Natural Language Processing and Language Modelling, *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, IEEE, pp.223–228.

[10] **Williams, R.J. and Zipser, D.** (1989). A Generalised Framework for Convolutional Decoding Using a Recurrent Neural Network, *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp.451–455.

[11] **Abdel-Hamid, O.**, **rahman Mohamed, A.**, **Jiang, H.**, **Deng, L.**, **Penn, G. and Yu, D.** (2014). Convolutional Neural Networks for Speech Recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545.

[12] **Lee, J.Y. and Dernoncourt, F.** (2016). Sequential Short-Text Classification with Recurrent and Convolutional Neural Networks, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, pp.515–520, `https://aclanthology.org/N16-1062/`.

[13] **Haykin, S. and Deng, C.** (1991). Classification of Radar Clutter Using Neural Networks, *IEEE Transactions on Neural Networks*, *2*(6), 589–600.

[14] **Grajal, J.**, **Quintas, A.G. and Lopez-Risueno, G.** (2005). MTD detector using convolutional neural networks, *IEEE International Radar Conference, 2005*, Arlington, VA, USA, pp.827–831.

[15] **Lopez-Risueno, G.**, **Grajal, J. and Diaz-Oliver, R.** (2003). Target detection in sea clutter using convolutional neural networks, *2003 IEEE Radar Conference (Cat. No.03CH37474)*, Huntsville, AL, USA, pp.321–328.

[16] **Wang, L.**, **Tang, J. and Liao, Q.** (2019). A Study on Radar Target Detection Based on Deep Neural Networks, *IEEE Sensors Letters*, *3*(3), 1–4.

[17] **Xie, Y.**, **Tang, J. and Wang, L.** (2019). Radar Target Detection Using Convolutional Neural Network in Clutter, *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, IEEE, pp.1–6.

[18] **Zhang, M.**, **Yang, Y.**, **Ji, Y.**, **Xie, N. and Shen, F.** (2018). Recurrent attention network using spatial-temporal relations for action recognition, *Signal Processing*, *145*, 137–145.

[19] **Ma, L.**, **Liu, M.**, **Wang, N.**, **Wang, L.**, **Yang, Y. and Wang, H.** (2020). Room-Level Fall Detection Based on Ultra-Wideband (UWB) Monostatic Radar and Convolutional Long Short-Term Memory (LSTM), *Sensors*, *20*(4), 1–17.

[20] **Tian, S.**, **Wang, W.**, **Ding, G. and Zhang, Z.** (2023). Target Detection in Sea Clutter with Transformer Neural Network, *Remote Sensing*, *15*(4), 1001.

[21] **Vaswani, A.**, **Shazeer, N.**, **Parmar, N.**, **Uszkoreit, J.**, **Jones, L.**, **Gomez, A.N.**, **Kaiser, and Polosukhin, I.** (2017). Attention is All You Need, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp.5998–6008.

[22] **Chen, H.**, **Qi, Z. and Shi, Z.** (2021). Remote sensing image change detection with transformers, *IEEE Geoscience and Remote Sensing Letters*, *18*(7), 1304–1308.

[23] **Koay, V.**, **Hong** *et al.* (2021). Shifted-Window Hierarchical Vision Transformer for Distracted Driver Detection, *Proceedings of TENSYMP*, pp.1–7.

[24] **Elbedwehy, S.** *et al.* (2022). Efficient Image Captioning Based on Vision Transformer Models, *CMC*, *73*(1), 1483–1500.

[25] **Dosovitskiy, A.**, **Beyer, L.**, **Kolesnikov, A.**, **Weissenborn, D.**, **Zhai, X.**, **Unterthiner, T.**, **Dehghani, M.**, **Minderer, M.**, **Heigold, G.**, **Gelly, S.**, **Uszkoreit, J. and Houlsby, N.** (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, *arXiv preprint arXiv:2010.11929*, https://arxiv.org/abs/2010.11929.

[26] **Liu, Z.** *et al.* (2021). Swin transformer: Hierarchical vision transformer using shifted windows, http://arxiv.org/abs/2103.14030, 2103.14030.

[27] **Gu, M.**, **Chen, Z.**, **Chen, K. and Pan, H.** (2023). IR-ST: A Lightweight transformer Network for Human Fall Detection Based on FMCW Radar, *IEEE Sensors Journal*, *PP*, 1–1.

[28] **Mahafza, B.R.** (2013). *Radar Systems Analysis and Design Using MATLAB*, CRC Press.

[29] **Electronics Desk** (yaklaşık 2019). *Radar System*, https://electronicsdesk.com/radar-system.html, erişim tarihi: 17 Haziran 2025.

[30] **Wang, X.**, **Li, J.**, **Yang, Y.**, **Lu, C.**, **Kwan, C. and Ayhan, B.** (2011). Comparison of Three Radar Systems for Through-the-Wall Sensing.

[31] **Hegazy, A.M.**, **Mosaad, M.M. and Hassan, A.M.** (2020). FMCW Software Defined Radar for Range and Speed Estimation, *Proceedings of the International Undergraduate Research Conference*, p. 29, https://www.researchgate.net/publication/305737436_FMCW_Software_Defined_Radar_for_Range_and_Speed_Estimation, accessed: 2025-06-17.

[32] **everything RF** (n.d.). *What Is Direct RF Sampling*, https://www.everythingrf.com/community/what-is-direct-rf-sampling, accessed: 2025-06-17.

[33] **Radar Tutorial** (n.d.). *Automatic Gain Control Methods*, https://www.radartutorial.eu/09.receivers/rx08.en.html, accessed: 17 June 2025.

[34] **Zheng, Q.**, **Jian, Y.**, **Wang, L.**, **Ma, Z.**, **Li, X.**, **Song, C.**, **Li, P. and Ding, L.** (2021). BPSK Modulation-Based Local Oscillator-Free IQ Demodulation for Millimeter Wave Imaging, *Journal of Sensors*, *2021*, 5596854, https://doi.org/10.1155/2021/5596854, accessed: 2025-06-17.

[35] **MathWorks** (n.d.). *Matched Filtering*, https://www.mathworks.com/help/phased/ug/matched-filtering.html, accessed: 2025-06-17.

[36] **Blunt, S.D. and Mokole, E.L.** (n.d.). *An Overview of Radar Waveform Diversity*, https://www.ittc.ku.edu/~sdblunt/papers/WD%20Tutorial%20Blunt%20Mokole%20-%20revised.pdf, accessed: 2025-06-17.

[37] **Abdelbagi, H.E.**, **Aljohani, M.**, **Mrebit, A. and Wicks, M.C.** (2015). FPGA-Based Coherent Doppler Processor for Marine Radar Applications, *Proceedings of the 2015 IEEE National Aerospace and Electronics Conference (NAECON)*, pp.1–6, https://www.researchgate.net/publication/283472549_FPGA-Based_Coherent_Doppler_Processor_for_Marine_Radar_Applications, accessed: 2025-06-17.

[38] **Access Engineering Library** (n.d.). *Homepage*, https://www.accessengineeringlibrary.com/, accessed: 2025-06-17.

[39] **Guillén Soriano, C.** (2020). Review of radar detectors with Constant False Alarm Rate, *Revista Telem@tica*, *19*, 78–90.

[40] **Coluccia, A.**, **Fascista, A. and Ricci, G.** (2020). CFAR Feature Plane: A Novel Framework for the Analysis and Design of Radar Detectors, *IEEE Transactions on Signal Processing*, *68*, 3903–3916.

[41] **Jalil, A.**, **Yousaf, H. and Baig, M.I.** (2015). Analysis of CFAR Techniques, *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, IEEE, Kuala Lumpur, Malaysia, pp.57–62.

[42] **Sahal, M.**, **Said, Z.A.**, **Putra, R.Y.**, **Kadir, R.E.A. and Firmansyah, A.A.** (2020). Comparison of CFAR Methods on Multiple Targets in Sea Clutter Using SPX-Radar-Simulator, *Proceedings of the 2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, IEEE, Surabaya, Indonesia, pp.260–265.

[43] **LeCun, Y.**, **Bengio, Y. and Hinton, G.** (2015). Deep learning, *Nature*, *521*(7553), 436–444.

[44] **O'Shea, K. and Nash, R.** (2015). An Introduction to Convolutional Neural Networks, *CoRR*, *abs/1511.08458*, http://arxiv.org/abs/1511.08458, 1511.08458.

[45] **Ergün, G.B. and Güney, S.** (2021). Classification of Canine Maturity and Bone Fracture Time Based on X-Ray Images of Long Bones, *IEEE Access*, *9*, 109004–109011.

[46] **Papa, L.**, **Russo, P.**, **Amerini, I. and Zhou, L.** (2023). A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking, *IEEE Transactions on Neural Networks and Learning Systems*, *34*(3), 1102–1128.

[47] **Kayacan, Y.E. and Erer, I.** (2024). A Vision-Transformer-Based Approach to Clutter Removal in GPR: DC-ViT, *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5, art no. 3505105.

[48] **Wu, H.**, **Yu, L.**, **Li, X.**, **Zhou, L.**, **Zhang, W. and Bai, G.** (2023). CTF-Net: A Convolutional and Transformer Fusion Network for SAR Ship Detection, *IEEE Geoscience and Remote Sensing Letters*, *20*, 4010005.

CURRICULUM VITAE

Name SURNAME: Sena ÇAYBAŞI

EDUCATION:

- **B.Sc.:** 2021, Istanbul Technical University, Electrical and Electronics Engineering, Electronics and Communications Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2021- Radar System Engineer at ASELSAN INC.

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Çaybaşı, S.**, Erer, I. (2025, May). Radar Target Detection using Improved Transformer Neural Network. *In 2025 33rd Signal Processing and Communications Applications Conference (SIU). (pp. 1-4) IEEE*