International Journal of Engineering Research and Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 22, Issue 8 (August 2026), PP 200-213

A Fine-Tuned Vision Transformer for Deepfake Image Detection

Kanwerjit
Dept.ofComputerScience&Engineering
PunjabiUniversity,Patiala,India

GauravDeep
Dept.ofComputerScience&Engineering
PunjabiUniversity,Patiala,India

Sumandeep Kaur
Dept.ofComputerScience&Engineering
PunjabiUniversity,Patiala,India

Abstract
The proliferation of algorithmically synthesized visual con-tent broadly labelled as deepfake poses a mounting threat to theintegrityofdigitalinformationecosystems. Despiterapid advancesingenerativemodelling, robustautomateddetection remains challenging, as synthesis quality now routinely ex-ceedsthethresholdofreliablehumaninspection.Thispaper fine-tunesaVisionTransformer(ViT)classifierfromthepub-licly released dima806/deepfake_vs_real_image_detection checkpoint via the Hugging Face Trainer API on a balanced corpus of 190,081 images drawn from Kaggle, comprising equalproportionsofauthenticphotographsandAI-generated samples spanning GAN-based and latent diffusion architec-tures.The fine-tuned model achieves an overall classifi-cation accuracy of 99.20% and a macro-averaged F1-score of 0.9920 on a held-out evaluation set of 38,081 images, withsymmetricper-classerrorrates(128falsepositives;175 false negatives).These results demonstrate that the Vision Transformer, whose globally unconstrained multi-head self-attentionmechanismenablesdetectionofthelong-rangespa-tial incoherence characteristic of synthetic imagery, consti-tutesacomputationallytractableandhigh-performingarchi-tectureforsynthetic-imageforensics,surpassingallsurveyed CNN and frequency-domain baselines on the same bench-mark.

Keywords—*deepfake detection; Vision Transformer (ViT); synthetic image classification; binary image forensics; fine-tuning; generative adversarial networks; latent diffusion models; transfer learning; Hugging Face.*

Date of Submission: 15-08-2026 Date of acceptance: 30-08-2026

I. INTRODUCTION

Deepfake, a portmanteau of *deep learning* and *fake*, entered commonusagein2017todescribeaclassofneurallysynthe-sized media in which generative models fabricate or manip-ulatefacialappearances,vocalsignatures,andbroaderscene contextswithsufficientfidelitytomisleadhumanobservers. Since that coinage, advances in generative modelling have consistentlyoutpacedthedevelopmentofreliableautomated detectiontools,yieldingapersistentasymmetrybetweensyn-thesiseaseanddetectiondifficulty.

Thesocial,political,and legal ramifications of this asymmetry are well documented: synthetic imagery has been exploited to manufacture polit-icaldisinformation,producenon-consensualintimatemate-rial, undermine evidentiary standards in legal proceedings, anderodepublictrustinphotographicrecordsastruthfulrep-resentations of reality [1,2].

The computational foundations of deepfake synthesis rest on unsupervised deep-learning frameworks that learn com-pact latent representations of visual domains and subse-quentlyexploitthoserepresentationstogenerateplausi-ble novel samples or to transform existing imagery in distribution-consistent ways.The encoder–decoder formu-lation, wherein an encoder maps an input image to a low-dimensionallatentcodeandacomplementarydecoderrecon-structsimageryfromthatcode,constitutesthebaseparadigm from which most contemporary synthesis pipelines are de-rived[2,3].Theintroductionofadversarialtraining,formal-ized by the Generative Adversarial Network (GAN) frame-work [4], substantially improved synthesis realism, while variational autoencoders [3] and score-based diffusion mod-els[5,6]furtherextendedgenerativeexpressivityandcondi-tional controllability.

TheVisionTransformer(ViT)[7]representsafundamen-tal departure from convolutional inductive biases in visual recognition.By decomposing input imagery into sequences offixed-

sizepatchembeddingsprocessedbyastandardtrans-former encoder with multi-head self-attention, ViT estab-lishesafullyglobalreceptivefieldfromtheveryfirstprocess-inglayer—anattributeinherentlysuitedtodetectingthelong-rangespatialincoherenceandcross-regionalsemanticviola-tions that frequently characterize AI-generated content.The present study empirically evaluates a fine-tuned ViT model appliedtobinarydeepfakeclassificationacrossalarge-scale balancedbenchmark,offeringatransparentandreproducible experimental record situated within the broader deepfake-detection literature [8,9].

A.ResearchObjectivesandScope

Threeprincipalresearchaimsgovernthisinvestigation.First, toundertakeacontrolledempiricalassessmentofViT-based binaryclassificationappliedtoalarge-scaledeepfakedetec-tion benchmark, documenting classification accuracy, per-class precision and recall, F1-score, and confusion-matrix statistics.Second,tocontextualizetheobservedperformance withintheextantdetectionliteratureviaastructuredmethod-ological comparativeanalysis.Third,toidentifythearchitec-tural properties of the Vision Transformer that are theoreti-callyandempiricallypertinenttothedeepfakedetectiontask, andtocharacterizethelimitationsandprospectiveextensions ofthepresentapproach[7,10].

II. RELATED WORK AND BACKGROUND

A. GenerativeArchitecturesforSyntheticImageProduction

Synthesizing deepfake content is fundamentally an exercise in learned distributional modelling:a generative network is optimized over a corpus of real imagery to internalize the statistical regularities of authentic visual data, and is subse-quently used to sample novel instances consistent with that learned distribution [2,3].The autoencoder constitutes the most elementary such architecture:an encoder E maps an input image x to a compact latent code $z=E(x)$, and a de-coder D reconstructs the image as $\hat{x}=D(z)$.The system istrainedtominimizethe reconstruction objective $\|x-\hat{x}\|$, ensuringthelearnedrepresentationencodestheperceptually significantcharacteristicsofthesourceimage. Post-training, decoding latent codes from alternative inputs enables face-attribute transfer across identities [3].

The GAN framework [4] remedied the blurring tendency ofvanillaautoencodersbycouplingthegenerativeobjective withanadversarialdiscriminatortrainedjointlywiththegen-erator to distinguish authentic from synthetic samples.The resulting adversarial equilibrium yields sharper outputs but riskstraininginstabilityandmodecollapse[11]. Progressive GAN [11] addressed this instability through a curriculum-styletrainingschemethatincrementallygrowsgeneratorand discriminator resolution.StyleGAN [12] introduced style-based modulation at each resolution scale, enabling fine-grained control over facial attributes while preserving glob-ally coherent photorealistic structure.These architectures dominated the face-synthesis landscape preceding the diffu-sion era and drove the majority of face-centric detection re-search from 2019 to 2022 [11].

Variational Autoencoders (VAEs) [3] offered a comple-mentaryprobabilisticgenerativeframework: theencoderpa-rameterizes a posterior distribution $q(z|x)$ over latentspace, enablingprincipledprobabilisticsampling.TheVAE objective combines a reconstruction term with a Kullback–Leibler divergence regularizer that constrains the posterior towardaGaussianprior,promotingdisentangledrepresenta-tions.Adversarial autoencoders substituted this regularizer with a latent-space adversarial discriminator, achieving im-proved sample quality relative to vanilla VAEs [13].

The most consequential recent advance in generative im-age modelling is the latent diffusion model [5,6], which iteratively denoises samples drawn from a Gaussian prior, guided by a score-matching objective that approximates the log-density gradient of the data distribution.Latent diffu-sion models achieve image quality surpassing GAN-based systems on perceptual benchmarks and offer superior mode coverage,controllability,andscalability.Critically,diffusion-synthesizedimagesexhibitfrequencyspectrathatcloselyap-proximatethoseofauthenticphotographs,underminingspec-tral detection heuristics effective against earlier GAN

out-puts [14,15].

B. DeepfakeDetectionMethodologies

Thedeepfakedetectionproblemiscanonicallyframedasa supervisedbinaryclassificationtask:adetectionfunction f mustassignalabel $y \in \{real, fake\}$ to each input image $x$withhighreliabilityand,ideally,withoutdependenceon generator-specificartefactsthatmaynotpersistacrossthe rapidlyevolving space of synthesis architectures [2,8].De-tectionmethodologyevolutionbroadlymirrorsthatofgener-ation, with successive detector generations motivated by the failuremodesofpredecessorswhenconfrontedwithmoreca-pable synthesis approaches.

The initial wave of detection research targeted physiolog-ical and geometric anomalies in facial synthesis—abnormal blink rates, unphysical gaze trajectories, or structurally in-consistent facial landmark geometries characteristic of early synthesis pipelines [16].While effective within constrained synthesisdistributions,theseapproachesprovedbrittletoim-provements in generation fidelity and failed to transfer to non-facial synthetic imagery.Forensic noise-fingerprinting methodsadaptedfromestablisheddigitalimageforensicsex-ploited characteristic noise residuals and compression sig-naturesassociatedwithspecificcamerasensorsandedit-ing software, but similarly failed to generalize as synthesis algorithms evolved to suppress these detectable signatures [17,18].

Deep CNN-based detection substantially elevated classi-ficationperformanceandcross-manipulationgeneralization. The compact MesoNet architecture [19] demonstrated com-petitiveface-swapdetectionthroughanend-to-endfour-layer CNN trained on balanced real/fake datasets.The Xception-baseddetectorevaluatedontheFaceForensics++benchmark [8]establishedthatlargeImageNet-pretrainednetworksfine-tuned for detection substantially outperform purpose-built shallow architectures, cementing transferlearning as the stan-dard methodology for subsequent CNN-based detection re-search.Wang et al. [17] further demonstrated that CNN de-tectors trained with appropriate augmentation on ProGAN outputscanachievestrongcross-generatorgeneralization,at-tributing this to systematic spectral distribution mismatches arisingfromtransposed-convolutionupsamplingcommonto CNN-based generators.

Spectral-domain analyses have supplied auxiliary detec-tion signals.GAN-generated images exhibit periodic arte-factsintheirdiscretecosinetransformspectraattributableto upsampling periodicity, and frequency-aware classifiers can exploit these cues [14].However, diffusion-based genera-tors bypass transposed-convolution upsampling entirely and therefore do not imprint the periodic spectral signatures that supportfrequency-domaindetection,motivatingtheshiftto-wardarchitecturessensitivetohigher-ordersemanticandspa-tial inconsistencies [15].

In this context, transformer-based detection architectures have demonstrated considerable promise.Their global modelling capability is directly relevant to detecting syn-thetic imagery that commonly exhibits locally plausible tex-turesandstructureswhileharboringgloballyinconsistent configurations—forexample,physicallyimplausibleillumi-nation gradients, incoherent depth-of-field renderings, or structurally anomalous proportional relationships requiring spatiallydistributedevidenceforreliableidentification[22, 23].

$Q_h = XW^Q, K_h = XW^K, V_h = XW^V$.Theattention outputforhead $h$is then:

$$\mathrm{Attn}_h(X) = \mathrm{softmax}_k\left(\frac{Q_h K^T}{\sqrt{d_h}}\right) V_h \quad (1)$$

III. VISIONTRANSFORMERARCHITECTURE

This section presents the theoretical foundations of the Vi-sion Transformer processing pipeline through which deep-fakeclassificationisperformed,fromrawpixelinputthrough patch tokenization, transformer encoding, and final binary prediction [7,24].

A. From Natural Language Processing to Visual Patch Se-quences

The transformer architecture [20], introduced for sequence-to-sequencelanguagemodelling,representedafundamen-tal departure from recurrent processing paradigms.Its cen-tral contribution was the replacement of sequential, state-dependent recurrence with parallel multi-head self-attention computedjointlyovertheentireinputsequence,enablingev-erytokentoattendtoeveryothersimultaneouslyandwithout positionalbias.Thisdesigneliminatesthevanishing-gradient pathologyinherenttodeeprecurrentchainsandpermitsprac-ticalscalingtosubstantiallylargermodelanddatasetregimes, providing the technical foundation for the pretraining-then-fine-tuning paradigm that characterizes modern NLP prac-tice [20,21].

Adapting this framework to visual inputs requires resolv-ingafundamentalrepresentationalincompatibility:language transformersoperateovercompact,discretetokensequences, whereas images are continuous, high-dimensional, and spa-tially structured signals.Prior attempts to incorporate self-attentionintovisionmodelsmanagedthequadraticcomplex-ityoffull-imageattentionbyconfiningcomputationstolocal spatialneighborhoods[23,25].Dosovitskiyetal.[7]circum-ventedthisconstraintthroughthepatchtokenizationstrategy thatdefinesViT:theinputimageispartitionedintoaregular, non-overlappinggridof$P \times P$ pixeltiles;eachtileisvector-izedandlinearlyprojectedintoa$d$-dimensionallatentspace; and the resulting sequence of patch tokens is processed by a standardtransformerencoderwithoutdomain-specificarchi-tectural modifications [7,26].This design choice preserves the full expressive capacity of the transformer while ren-der-ingtheimagecompatiblewithsequence-basedprocessingat manageable computational cost.

B. Multi-HeadSelf-AttentionandPositionalEncoding

Thecorecomputationalprimitiveofthetransformerencoder ismulti-headself-attention.Given$N$inputtokenembeddings arranged as $\mathbf{X} \in \mathbb{R}^{N \times d}$, each attention head h constructsquery,key,andvalueprojectionsthroughlearnedweightma-

Outputs from all H heads are concatenated and projected byalearnedlinearmaptoproducethemulti-headoutput. Scalingtheinnerproductby$\sqrt{d_k}$preventssoftmaxsatura-tion and thus gradient vanishing at large feature dimension-alities[20].Critically,theattentionoperationiscomputed jointlyoverall$N \times N$tokenpairs,establishingaglobally un-constrainedreceptivefieldinwhicheverypatchtokencandi-rectlyconditioneveryothertoken'srepresentationateachen-coderdepth,irrespectiveofgeometricseparationintheorig-inal image [7,24,27].

Because self-attention is inherently permutation-equivariant—producing identical outputs for any reordering oftheinputsequence—spatiallayoutinformationmustbe supplied explicitly.ViT employs learnable positional embeddings:asetof$d$-dimensionalparametervectors$\{p_1, p_2 \ldots, p_N\}$one per spatial position, which are added element-wise to the corresponding patch embeddings prior to encoder ingestion.These vectors are optimized jointly withallothermodelparametersduringpretraining,allowing the network to acquire spatial relational structure empiri-cally suited to the training distribution, rather than being constrained to hand-designed sinusoidal bases [7,28].

C. ClassificationTokenandEncoderArchitecture

Following the BERT sequence classification convention [21], ViTprependsalearnableclassificationtoken, denoted[CLS], tothepatch-embeddingsequencebeforeencoderprocessing. This token is not associated with any image region; rather, its final-layer representation serves as the holistic image-level summary passed to the downstream task head.Across the encoder's L layers, the [CLS] token attends bidirectionallyto all patch tokens,

progressively accumulating global im-age evidence.Its terminal representation is routed througha single linear projection to produce class logits during fine-tuning [7,21,24].

Each encoder layer consists of two sequentially stacked sub-modules:a multi-head self-attention sublayer and a position-wise feed-forward network (FFN) sublayer.Pre-layer normalization is applied before each sub-module, and residual skip connections bypass each, together promoting stable gradient propagation through deep stacks [29].The FFN applies a two-layer projection with GELU activation:

$$\mathrm{FFN}(x)=W^2\cdot\mathrm{GELU}W^1x+b^1+b^2 \quad (2)$$

wherethehiddendimensionalityissetto$4d$. Thealternation ofgloballymixedattentionandposition-wisenon-linearpro-jectionisiterated$L$ times,progressivelytransformingpatch tricesW_h^Q, W_h^K, W_h^V ∈ R^{d×d_k},where$d_k=d/H$denotes representationstowardthetask-discriminativefeaturesaggre-theper-headdimensionalityforan$H$-headconfiguration: gatedinthefinal[CLS]embedding[7].

D. RelevancetoDeepfakeDetection

Several architectural properties of ViT motivate its applica-tion to deepfake forensics on principled theoretical grounds [9,22].Thegloballyunconstrainedattentionmechanismen-ablesthemodeltoidentifycoherenceviolationsbetweenspa-tiallynon-adjacentimageregions—forexample,illumination gradients traversing a face that conflict with ambient scene lighting, or facial geometric proportions that appear locally plausiblebutaregloballyinconsistentwiththestructuralreg-ularitiesofauthentichumananatomy.Detectingsuchcross-region anomalies requires integrating evidence from distant spatialpositions,acapabilitythatconvolutionalarchitectures approach only through very deep stacks with correspondingly large effective receptive fields [30,31].

Complementing this global reasoning capacity, ViT's patch-basedprocessingavoidstheprogressivespatialdown-sampling intrinsic to CNN pooling hierarchies, which trade fine-grained spatial resolution for increasing representa-tionalabstraction[32].Bydeferringallspatialaggrega-tion to the [CLS] token pooling step and preserving fullper-patch resolution through all encoder layers, ViT retains localized information that may encode synthesis-specific residuals—including texture discontinuities at patch bound-aries, locally anomalous spectral statistics, and spatially structured noise patterns characteristic of particular gener-ator implementations—that would be irreversibly discarded by earlier pooling stages in a conventional convolutional pipeline.Together,thesepropertiesmakeViTarchitecturally well-suited to the dual demands of deepfake detection:sen-sitivity to subtle, spatially localized generation artefacts and capacityforsimultaneousglobalimagecoherencereasoning.

IV. METHODOLOGY

Theproposeddetectionsystemimplementsafifteen-stepend-to-end pipeline, summarized in Table I and illustrated inFig. 1, transforming a raw input image into a binary authen-ticity decision.The pipeline covers dataset acquisition and stratified partitioning; image standardization to $224\times224$and channel normalization using ImageNet statistics ($\mu=[0.485,0.456,0.406]$,$\sigma=[0.229,0.224,0.225]$);training-onlystochasticaugmentation;patchtokenizationinto$N=196$ non-overlapping16×16 tiles; linearpatchprojection intothe$d=768$latentspace;additivelearnablepositional embeddingsandaprepended[CLS]token(yielding197to-kens); processing through $L=12$ transformer encoder lay-ers with 12-head self-attention; [CLS]-token aggregation; a linear classification head producing logits $[z_{\text{real}},z_{\text{fake}}]$; fine-tuning with AdamW; inference on the held-out partition; argmaxdecoding;andfinalmetriccomputationfromthecon-fusion matrix.

A. TrainingConfigurationandHyperparameters

All experiments were conducted on a Kaggle-hosted envi-ronment equipped with a single NVIDIA T4 GPU (16 GB GDDR6),running Python 3.10,PyTorch 2.1.0,CUDA 12.1,

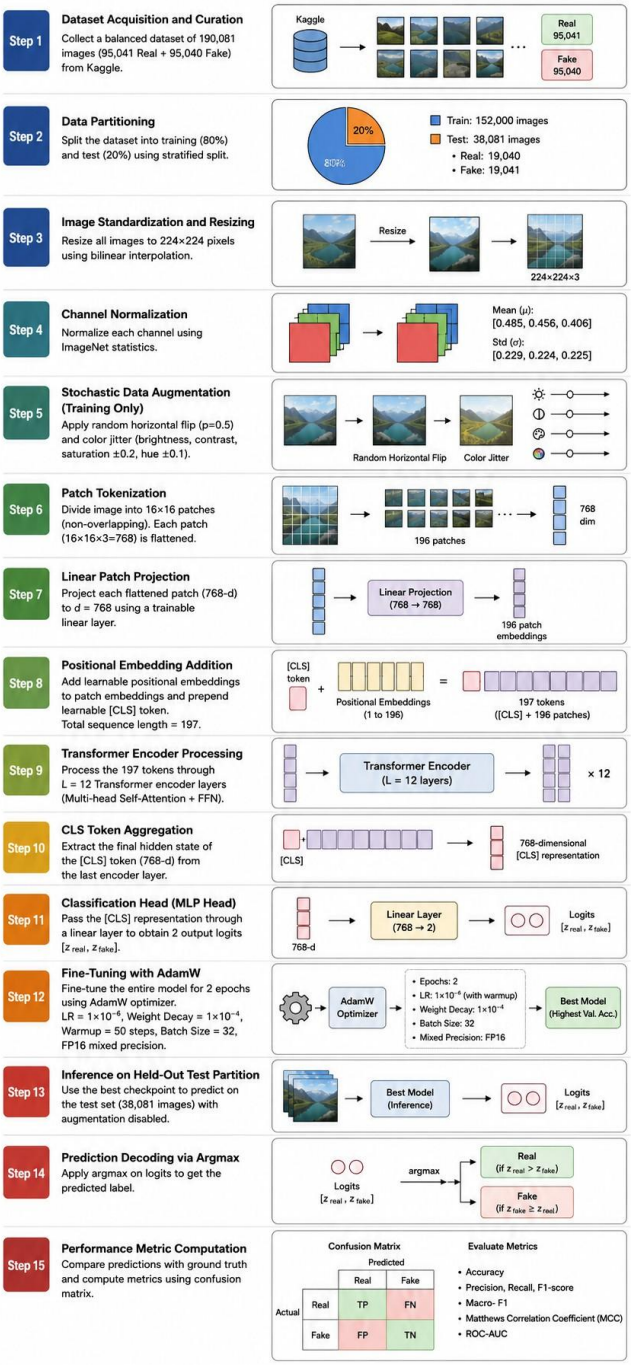

Fig.1.WorkingoftheViTmodel:thefifteen-stepend-to-enddetectionpipeline.

TABLEI
Fifteen-StepEnd-to-EndViTDetectionPipeline

TABLEII
TrainingHyperparameterSpecification(ViT-Base/16)

Step	Stage
1	Datasetacquisitionandcuration(190,081images)
2	Stratified80:20partitioning(152,000/38,081)
3	Imagestandardizationandresizing(224×224)
4	Channelnormalization(ImageNetstatistics)
5	Stochasticdataaugmentation(trainingonly)
6	Patchtokenization(N=196tokens)
7	Linearpatchprojection(d=768)
8	Positionalembedding+[CLS]token(197tokens)
9	Transformerencoderprocessing(L=12)
10	[CLS]tokenaggregation
11	Classificationhead(MLPhead)
12	Fine-tuningwithAdamW(2epochs)
13	Inferenceonheld-outtestpartition
14	Predictiondecodingviaargmax
15	Performancemetriccomputation

Hyperparameter	Value
Trainingepochs	2
Initiallearningrate	1×10^{-5}
Optimizer	AdamW
Weightdecay	1×10^{-4}
LRwarmupsteps	50(linear)
Trainingbatchsize	32
Evaluationbatchsize	8
Mixedprecision	FP16(AMP)
Randomseed	42
GPU	NVIDIAT4(16GBGDDR6)
Framework	PyTorch2.1.0/HFTransformers4.40.0

Recall measures the proportion of actual positive instances correctly detected:

andHuggingFaceTransformers4.40.0.Therandomseed wasfixedto42acrossPyTorch,NumPy,andPython'sbuilt-

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{5}$$

inrandommodulepriortodatasetpartitioningandmodelini-tialization to ensure full experimental reproducibility.Ta-ble II summarizes the complete hyperparameter specification.AdamWwasselectedoverstandardAdamforitsde-

Specificitymeasurestheproportionofactualnegativein-stances correctly classified:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

coupledweightdecayformulation,whichappliesregular-izationdirectlytoparametermagnitudesratherthanfold-

ing it into the adaptive gradient scaling—a property demon-stratedtoimprovefine-tuninggeneralizationforlargepre-

TheF1-scoreistheharmonicmeanofprecisionandrecallfor class c:

trainedmodels[10].Theconservativelylowinitiallearn-ingrateof1×10^{-6}wasdeliberatelychosentolimitcatas-

$$\text{F1}_c = 2\times\frac{\text{Precision}_c\times\text{Recall}_c}{\text{Precision}_c+\text{Recall}_c} \tag{7}$$

trophicforgetting:thedegradationofpretrainedrepresenta-

tional structure that occurs when fine-tuning step sizes substantiallyexceedtheeffectiveoptimizationregimeofthebase checkpoint [35].A 50-step linear warmup ramp was appliedtoallowAdamW'sfirst-andsecond-ordermomentac-cumulators to stabilize before the full learning rate became active.The two-epoch training duration,while brief rela-tive to training-from-scratch regimes, is consistent with the accelerated convergence behavior characteristic of transfer learning from checkpoints already well-aligned with the tar-get task distribution, and the Hugging Face Trainer API's `load_best_model_at_end`mechanismguaranteesthat the deployed model corresponds to the highest validation-accuracy checkpoint.

B. MetricDefinitions

All evaluation metrics are derived from the four fundamen-tal counts of the 22 confusion matrix. Overall accuracy measures the proportion of correctly classified instances:

The symmetric macro-averaged F1-score of 0.9920 across both classes indicates equal discriminative reliability, without systematic favoritism toward either category.

V. EXPERIMENTALRESULTS

A. Dataset

Experimental evaluation was conducted on a large-scale benchmarkcorpusof190,081imagesretrievedfromtheKag-gleplatform,comprisingabalancedmixtureofauthenticpho-tographs and AI-generated synthetic images drawn from di-versegenerativearchitectures,includingGAN-basedandla-tent diffusion models.The deliberate class equilibrium—equal instance counts for both categories—ensures that standard classification metrics, including overall accuracy, macro-averaged F1-score, and per-class precision and re-call, remain directly interpretable without prior-probability correction and preventsthe optimizer from exploiting triv-ialmajority-classassignmentheuristics[8,33].Thecom-

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

plete corpus was partitioned into training and held-out evaluationsubsetsusingastratified80:20allocationviascikit-

Precisionmeasurestheproportionofpositivepredictionsthat are genuinely positive, per class c:

learn1.4.0(`train_test_split`,shuffle=True),yielding 152,000 training images and a fixed test partition of 38,081 images(19,040authentic;19,041synthetic).Thetestpar-

$$\text{Precision}_c \equiv \frac{TP_c}{TP_c + FP_c} \tag{4}$$

tition was isolated prior to any model training and was not exposed to the training pipeline at any stage.

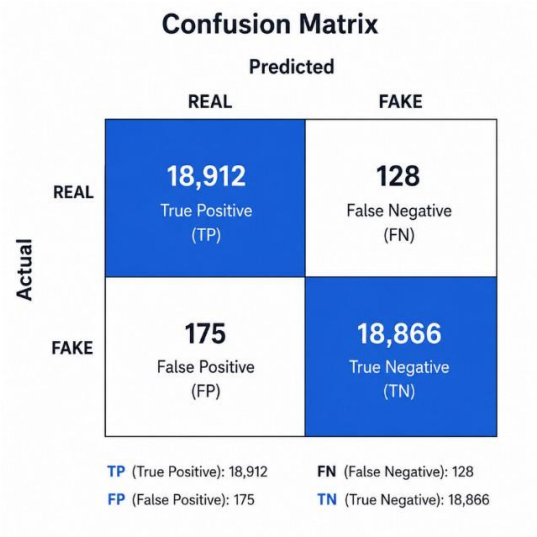

Fig. 2.Confusion matrix of the fine-tuned ViT-Base/16 model on the held-outtestpartition. Cellcountsareauthoritative;seeTableIIIforthelabelled counts ($FP = 128$, $FN = 175$).

TABLEIII
ConfusionMatrix(ViT-Base/16,N=38,081)

Pred:Real		**Pred:Fake**	**Total**
Actual:Real	18,912(TP)	175 (FN)	19,040
Actual:Fake	128 (FP)	18,866(TN)	19,041
Total	19,040	19,041	38,081

B. ConfusionMatrixAnalysis

The confusion matrix (Fig. 2, Table III) records $TP = 18{,}912$correctlyidentifiedauthenticimages,$FN = 175$ authentic images misclassified as synthetic, $FP = 128$ synthetic images misclassified as authentic, and $TN = 18{,}866$correctlyidentifiedsyntheticimages.Thenear-symmetricerrordistribution—128falsepositivesagainst175 false negatives—confirms the absence of systematic directionalbias,adesirablepropertyinbalanceddeployment contexts where false-alarm and missed-detection costs areof comparable magnitude.The aggregate misclassification countof303across38,081testimagescorrespondstoaner-ror rate of 0.80%, fully consistent with the reported 99.20% accuracy.

C. Per-ClassClassificationReport

Ofallimagesassigneda*Real*prediction,$18{,}912/(18{,}912 + 128) = 0.9933$weregenuinelyauthentic;ofallimagesassignedа*Fake*prediction,$18{,}866/(18{,}866+175) = 0.9908$ weregenuinelysynthetic. Thenear-symmetricperclasspre-cisionandrecallconfirmbalanceddiscriminationacrossboth categories, consistent with training on a class-

balanced cor-pus.

TABLEIV
Per-ClassClassificationReport(N=38,081)

Class	Prec.	Recall	F1	Support
Real	0.9933	0.9908	0.9921	19,040
Fake	0.9908	0.9933	0.9920	19,041
MacroAvg.	0.9920	0.9920	0.9920	38,081
Accuracy		99.20%		38,081

D. ComparativeEvaluation

The proposed ViT-Base/16 achieves the highest accuracy (99.20%) and F1-score (0.9920) across all surveyed meth-ods(TableV).Amongarchitecturesevaluatedoncomparable Kaggle corpora, the nearest competitor is the EfficientNet-B0convolutionalbaseline(98.45%),followedbyUniver-sal Fake Image Detection (Ojha et al. [37], 96.10%).The Multi-Attentional Deepfake Detection system [38] achieves 97.60%, but on the FaceForensics++ facial manipulation benchmark—a more constrained setting than the mixed GAN-and-diffusioncorpususedhere,whichrendersthatfig-ure a narrower measure of general-purpose detection capa-bility.First-generation approaches such as MesoNet [19] (83.00%) and Freq-CNN [14] (91.80%) exhibit substan-tially lower accuracy, consistent with their design targeting spectral periodicity artefacts characteristic of pre-diffusion GAN synthesis—signatures that are largely absent in mod-erndiffusion-modeloutputsconstitutingasignificantportion oftheevaluationcorpus.TheproposedViT-basedapproach surpasses MesoNetby 16.20 percentage points, amargin re-flectingbotharchitecturaladvancementandthesubstantially larger, more diverse training corpus employed in this work.

VI. LIMITATIONS

Thepresentstudyissubjecttoseveralmethodologicalandap-pliedconstraintsthatconditionthegeneralisabilityandcom-pleteness of the reported findings [39].

*Trainingdatatemporalcoverage:*Thetrainingcorpus, whilstlargerelativetomanypublisheddeepfakedetection benchmarks,represents a snapshot of the generative model landscape at a fixed point in time.The model has not been assessed against synthetic outputs from generative architec-tures released subsequent to the data collection window, in-cludingmorerecentdiffusionvariants,videoframesynthesis systems,andhybrideditingworkflows.Giventhecontinuous pace of advancement in generative modelling, this temporal coveragegapwidensovertime, potentiallyunderminingde-tectionperformanceagainstthelatestsynthesismethods[40].*Restrictedtrainingduration:*Thetwo-epochfine-tuning schedulewasimposedbythecomputationalbudgetofthe KaggleT4environmentratherthanbyprincipledablationof theoptimalstoppingpoint. Additionaltrainingepochs,com-bined with learning-rate schedule adjustments, could plausi-blyyieldfurtherperformancegains;accordingly,thepresent results should be interpreted as a conservative lower bound onachievableaccuracyunderunconstrainedtrainingcondi-

TABLEV
ConsolidatedPerformanceComparisonAcrossMethods

Model	Year	Dataset	Accuracy	Precision	Recall	F1-Score	Ref.
MesoNet	2018	FF++	83.00	0.831	0.829	0.830	[19]
XceptionNet	2019	FF++	95.73	0.961	0.953	0.957	[8]
CNN-F(Wangetal.)	2020	KaggleAI-Gen.	92.40	0.927	0.921	0.924	[17]
Freq-CNN(Franketal.)	2020	GAN/Diff. Mixed	91.80	0.920	0.916	0.918	[14]
Multi-Att.(Zhaoetal.)	2021	FF++	97.60	0.978	0.974	0.976	[38]
UniversalFID(Ojhaetal.)	2023	KaggleAI-Gen.	96.10	0.963	0.960	0.961	[37]
EfficientNet-B0(baseline)	2025	Kaggle(190K)	98.45	0.9845	0.9845	0.9845	—
ViT-Base/16(proposed)	**2025**	**Kaggle(190K)**	**99.20**	**0.9920**	**0.9920**	**0.9920**	—

FF++ = FaceForensics++. Kaggle (190K) refers to the balanced 190,081-image corpus used in this work. External results are as reported in the citedpublications;differencesindatasetscope,corpussize,andevaluationprotocolprecludestrictnumericalequivalenceacrossrows. TheEfficientNet-B0rowisa convolutional baseline evaluated on the same corpus and is reported for reference.

tions[35]. Futureworkshouldconductsystematicepochand learning-rate ablations.

*Absence of spatial localisation:*The binary global classi-ficationobjectiveyieldsnospatialindicationofwhichimage regions drive the model's synthetic-content verdict.Foren-sicusecasesrequirenotmerelyanimage-levelauthentic-ity decision but also spatially localised evidence of anoma-lous content—a capability requiring dense prediction archi-tectures,gradient-basedattributiontechniquessuchasGrad-CAM,orattention-mapanalysistoolsnotincorporatedinthe present experimental design [31].

*Post-processing and adversarial robustness:*Evaluation images were presented unmodified.In operational deploy-mentcontexts,imagesroutinelyundergoJPEGcompression, spatial rescaling, colour correction, watermark embedding, platform re-encoding, or adversarial perturbations targeting detection evasion.Performance under these conditions has not been characterised and is expected to degrade substan-tially for heavily post-processed or adversarially perturbed inputs [30].

VII. FUTURE WORKANDRESEARCHDIRECTIONS

A. Large-ScaleandTemporallyUpdatedTraining

The most immediate avenue for extending the present work involves systematic evaluation of detection efficacy as a function of training corpus size and temporal breadth.Fu-ture investigations should assemble training datasets span-ning synthetic imagery from a broader range of generative architectures—encompassingthelatestdiffusion-basedmod-els,videosynthesisplatforms,andhybrideditingpipelines—tocharacterisethescalingbehaviourofViT-baseddetec-tion and to quantify the rate at which performance degrades as temporal distance between training-data provenance and deployment-era synthesis tools increases [40].The incor-porationofcontinuallearningmechanisms,includingelastic weightconsolidation,experiencereplayovercuratedbuffers of past generator outputs, and parameter-efficient adapter modules,wouldequipdetectionsystemstoassimilatenewly emerging synthesis techniques without complete model re-training.

B. MultimodalandCross-ModalDetection

Thecurrentworkconfinedanalysistostandalonestill-image classification, setting aside contextual signals—associated metadata, natural language captions, social network prove-nance, or temporal video information—available in realistic deployment environments.Integrating textual or contextual signalsthroughcross-modalattentionmechanismsorvision-languagearchitecturescouldsupplycomplementarydiscrim-inative cues, particularly for identifying synthetic images whosepixel-levelcharacteristicsappearlocallyplausiblebut whosesemanticcontentiscontextuallyanomalousrelativeto accompanying text [2].Extending the framework to video modalities, with explicit modelling of temporal consistency across consecutive frames, would directly address the rising prevalence of deepfake video content [16].

C. ModelCompressionforEfficientDeployment

The ViT-Base model's approximately 86 million parame-ters impose a non-negligible inference overhead for high-volumecontentmoderationpipelines.Systematicstudy of quantisation-aware training, structured magnitude prun-ing,andknowledgedistillationintosmallerstudentarchitec-tures should be pursued to establish the accuracy-efficiency trade-off frontier for this task [34,36].Fine-tuning smaller ViT variants—including ViT-Tiny and ViT-Small—on the same benchmark would enable controlled comparisons of accuracy-latency trade-offs across model scales [7,32].

D. ExplainabilityandSpatialAttribution

Addressing the spatial localisation limitation identified in Section VI, future work should integrate attention rollout, Grad-CAM, or transformer-specific attribution methods to produce saliency maps highlighting image regions that most strongly influence the model's real/fake verdict.Such visu-alisations would enhance forensic interpretability, facilitate erroranalysisformisclassifiedsamples,andsupportdomain expertsinvalidatingmodeldecisions—aprerequisiteforde-ploymentinhigh-stakesforensic,journalistic,andlegalcon-texts [31].

VIII. CONCLUSION

This paper has presented a rigorous empirical evaluation of Vision Transformer-based binary classification for the iden-tificationofAI-generateddeepfakeimagery,establishingthe architecture as a high-performing and computationally prac-tical solution for this forensically critical task.Fine-tuning the*dima806/deepfake_vs_real_image_detection*ViTcheck-pointovertwooptimisationepochsonabalancedlarge-scale dataset of 190,081 images yielded a test-set classification accuracy of 99.20%, a macro-averaged F1-score of 0.9920, near-symmetric per-class precision and recall, and a total of only 303 misclassifications across 38,081 test images [7,8].

The theoretical grounding of these results in ViT's ar-chitectural properties was examined comprehensively:the model's globally unconstrained receptive field via multi-head self-attentionenablesdetectionoflong-rangespatialincoher-encepatternsinaccessibletoconvolutionalnetworkswithout verydeepstacking;thepatch-basedtokenrepresentationsus-tains full spatial resolution through all encoder layers, pre-serving sensitivity to localised synthesis artefacts while en-ablingholisticevidenceaggregationthroughthe[CLS]pool-ing mechanism [7,24]; and the pre-norm transformer en-coder structure promotes stable gradient propagation, facil-itatingreliablefine-tuningtowardforensicdiscriminationob-jectives.The results were contextualised within a struc-tured review of the deepfake generation and detection liter-ature, and comparison with published benchmarks demon-stratesthatthepresentfine-tunedViTachievesperformance competitive with or superior to recently proposed detection systems while remaining tractable within the resource con-straints of a single consumer-grade GPU. The limitations identified—encompassing training data temporal coverage, spatial localisation, post-processing robustness, and adver-sarialvulnerability—collectivelydelineatearesearchagenda whosesystematicpursuitwillbenecessarytobridgethegap between the strong in-distribution performance documented hereandtherobust, interpretable,andoperationallyscalable detection systems required for real-world deployment.

ACKNOWLEDGMENTS

TheauthorsthankthecreatorsoftheKaggledeepfakedetectiondatasetandthecontributorstothe Hugging Face Transformers library and the *dima806/deepfake_vs_real_image_detection* model check-point.Computational resources were provided via Kaggle's free GPU tier.

DATAAVAILABILITY

ThedatasetusedinthisstudyispubliclyavailableontheKag-gle platform.The fine-tuned model checkpoint and evalua-tionscriptswillbemadeavailableuponreasonablerequestto the corresponding author.

REFERENCES

[1] B. Chesney and D. Citron, "Deep fakes:A looming challenge forprivacy,democracy,andnationalsecurity,"*Calif.LawRev.*,vol.107,no. 6, pp. 1753–1820, 2019.
[2] R.Tolosana, R.Vera-Rodriguez, J.Fierrez, A.Morales, andJ.Ortega-Garcia,"DeepFakesandBeyond:ASurveyofFaceManipulationandFake Detection," *arXiv:2001.00179*, 2020.
[3] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes,"*arXiv:1312.6114*,2014.
[4] I. J. Goodfellow et al., "Generative Adversarial Networks,"*arXiv:1406.2661*,2014.
[5] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion ProbabilisticModels," *arXiv:2006.11239*, 2020.
[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer,"High-Resolution Image Synthesis with Latent Diffusion Models,"*arXiv:2112.10752*,2022.
[7] A.Dosovitskiyetal.,"AnImageisWorth16x16Words:Transformersfor Image Recognition at Scale," *arXiv:2010.11929*, 2021.
[8] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M.Nießner, "FaceForensics++:Learning to Detect Manipulated FacialImages," *arXiv:1901.08971*, 2019.
[9] D.Gragnaniello,D.Cozzolino,F.Marra,G.Poggi,andL.Verdoliva,"AreGANgeneratedimageseasytodetect? Acriticalanalysisofthestate-of-the-art," *arXiv:2104.02617*, 2021.
[10] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regulariza-tion," *arXiv:1711.05101*, 2019.
[11] T.Karras,T.Aila,S.Laine,andJ.Lehtinen,"ProgressiveGrowingofGANsforImprovedQuality,Stability,andVariation,"in*Proc.ICLR*,2018.
[12] T.Karras,S.Laine,andT.Aila,"AStyle-BasedGeneratorArchitec-tureforGenerativeAdversarialNetworks,"*arXiv:1812.04948*,2019.
[13] P.Isola,J.-Y.Zhu,T.Zhou,andA.A.Efros,"Image-to-ImageTrans-lation with Conditional Adversarial Networks," *arXiv:1611.07004*,2018.
[14] J.Frank,T.Eisenhofer,L.Schönherr,A.Fischer,D.Kolossa,andT.Holz,"LeveragingFrequencyAnalysisforDeepFakeImageRecog-nition," *arXiv:2003.08685*, 2020.
[15] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L.Verdoliva, "Onthedetectionofsyntheticimagesgeneratedbydiffu-sion models," *arXiv:2211.00680*, 2022.
[16] Y.Li,M.-C.Chang,andS.Lyu,"InIctuOculi: ExposingAICreatedFakeVideosbyDetectingEyeBlinking,"in*Proc.IEEEWIFS*,2018.
[17] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros,"CNN-generatedimages are surprisingly easy to spot...for now,"*arXiv:1912.11035*,2020.
[18] L.Verdoliva,"MediaForensicsandDeepFakes:anoverview,"*IEEE*

J.Sel.TopicsSignalProcess.,2020.
[19] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet:aCompact Facial Video Forgery Detection Network," in *Proc. IEEEWIFS*,2018.
[20] A. Vaswani et al., "Attention Is All You Need," in *Adv. Neural Inf.Process. Syst.*, 2017.
[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019.
[22] A.Haliassos,K.Vougioukas,S.Petridis,andM.Pantic,"LipsDon'tLie:A Generalisable and Robust Approach to Face Forgery Detection," in *Proc. IEEE/CVF CVPR*, 2021.
[23] N.Parmaretal.,"ImageTransformer,"*arXiv:1802.05751*,2018.
[24] T.Wolfetal.,"HuggingFace'sTransformers:State-of-the-artNaturalLanguage Processing," *arXiv:1910.03771*, 2020.
[25] P.Ramachandran,N.Parmar,A.Vaswani,I.Bello,A.Levskaya,and
J.Shlens,"Stand-AloneSelf-AttentioninVisionModels,"in*Adv.Neural Inf. Process. Syst.*, 2019.
[26] H.Touvron,M.Cord,M.Douze,F.Massa,A.Sablayrolles,and
H. Jégou,"Training data-efficient image transformers & distillationthrough attention," in *Proc. ICML*, 2021.
[27] Z. Liu et al., "Swin Transformer:Hierarchical Vision Transformerusing Shifted Windows," in *Proc. IEEE/CVF ICCV*, 2021.
[28] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling VisionTransformers," *arXiv:2106.04560*, 2022.
[29] R. Xiong et al., "On Layer Normalization in the Transformer Archi-tecture," *arXiv:2002.04745*, 2020.
[30] N.CarliniandH.Farid,"EvadingDeepfake-ImageDetectorswithWhite- and Black-Box Attacks," in *Proc. IEEE/CVF CVPRW*, 2020.
[31] P.Zhou,X.Han,V.I.Morariu,andL.S.Davis,"Two-StreamNeuralNetworks for Tampered Face Detection," *arXiv:1803.11276*, 2018.
[32] M. Tan and Q. V. Le, "EfficientNet:Rethinking Model Scaling forConvolutional Neural Networks," *arXiv:1905.11946*, 2019.
[33] Y.MirskyandW.Lee,"TheCreationandDetectionofDeepfakes: ASurvey," *ACM Comput. Surv.*, 2020.
[34] C.Sun, A.Shrivastava, S.Singh, andA.Gupta, "RevisitingUnreasonableEffectivenessofDatainDeepLearningEra,"*arXiv:1707.02968*,2017.
[35] E. L. Aleixo, J. G. Colonna, M. Cristo, and E. Fernandes, "Catas-trophicForgettinginDeepLearning: AComprehensiveTaxonomy,"*arXiv preprint*, 2023.
[36] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in aNeural Network," *arXiv:1503.02531*, 2015.
[37] U.Ojha,Y.Li,andY.J.Lee,"TowardsUniversalFakeImageDetec-tors that Generalize Across Generative Models," *arXiv:2302.10174*,2023.
[38] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional Deepfake Detection," in *Proc. IEEE/CVF CVPR*, 2021.
[39] C. Szegedy et al., "Intriguing properties of neural networks,"*arXiv:1312.6199*,2014.
[40] H.Dang,F.Liu,J.Stehouwer,X.Liu,andA.Jain,"OntheDetectionof Digital Face Manipulation," *arXiv:1910.01717*, 2020.