End To End Speech Emotion Recognition Using CNN And Spectrogram with Gender Prediction

Hemalatha B[1], Dr Krishna Kumar P R[2]
[1]*M. Tech Student, Department of CSE & Technology, SEA College of Engineering & Technology*
[2]*Professor and HOD Department of CSE, SEA College of Engineering & Technology, Bangalore*

***Abstract*—Figuring out feelings through voice has become a big deal in AI and machine learning circles. This effort aims to build a smart setup that catches emotions in spoken words - at the same time guessing if the speaker is male or female - by relying on convolutional neural networks alongside sound visuals called spectrograms. Instead of treating sounds as raw data, it turns them into images first. Working behind the scenes, signal processing joins forces with deep learning models so emotional tones can be sorted more precisely. Accuracy steps up when both types of tech run together rather than alone.**

From open-source voice collections showing feelings like joy, sorrow, rage, terror, shock, dislike, and no emotion at all - sound files are gathered here. Before anything else happens, background hums get stripped out; clarity takes priority during cleanup steps. Normalization adjusts volume peaks, while quiet gaps vanish so every clip holds useful information only. Changing playback speed uniformly helps align inputs across different sources despite original variations. Once cleaned, key characteristics emerge through image-like maps built from sound waves over time. Mel-scale pictures of frequency shifts reveal how voices twist when emotions change unexpectedly. MFCC values pull out subtle vocal textures machines later recognize as distinct mood markers. These visuals act much like snapshots that highlight rhythm, pitch bends, and intensity bursts uniquely tied to each feeling. Convolution networks study those images just as they would photos - spotting tiny clues hidden in wave shapes. Emotion-linked habits within speech slowly become visible after repeated scanning by learning systems.

From those captured spectrogram pictures, a special kind of neural network takes over - its job is spotting patterns tied to emotion and voice differences between genders. Instead of hand-picking features, the system learns them on its own through layered processing steps. Layer after layer, it sharpens what it recognizes, stacking convolutions with down sampling and fully connected stages. Performance gets checked thoroughly, using measures like correct guess rates, consistency scores, detection completeness, balance indicators, plus detailed error mapping. Each step feeds into how well the whole setup classifies voices.

Accuracy matters most when spotting feelings and voices in sound, yet speed cannot slow down. One idea runs through every use: machines that listen better start here. Virtual helpers answer questions, sure - but now they might sense stress too. Customer service tools shift quietly, adapting to tone instead of just words. Mental health check ins? They gain subtle clues from how someone speaks, not only what is said. Call centers grow smarter by catching frustration before it spreads. Chatbots react differently if sadness shows up mid-sentence. Security setups watch for vocal tension like a guard watching shadows. Even classrooms change when software adjusts lessons based on student mood shifts. Deep networks meet audio filters somewhere between labs and daily life. Human talk gets broken into pieces, then rebuilt as meaning machines grasp. Interaction evolves without flashy promises - simply closer mimicry of natural listening. Understanding grows not from rules, but patterns found in thousands of spoken moments.

I. INTRODUCTION

Most people find talking the easiest way to share thoughts. Besides passing on facts, spoken words often show how someone feels inside. Feelings like joy, sorrow, rage, panic, shock, dislike, or nothing at all come out in shifts in tone, highness or lowness of voice, timing, loudness, and sound waves. Machines learning to pick up on these cues now draw serious attention in today's tech studies. Computers can now sense feelings in how people speak, thanks to tools like artificial intelligence and deep learning. Because of this, machines understand us better when we talk, which makes conversations flow easier. Instead of

just hearing words, they pick up on emotional cues hidden in voice patterns. Systems built this way respond in ways that feel less robotic, more tuned in. As progress continues, talking to devices begins to resemble real human exchanges.

Lately, computers that detect feelings in voices have become more popular. Fast changes in artificial intelligence plus how people interact with machines helped push this forward. Old voice tools focused on turning sounds into written words, yet missed whether someone sounded happy or upset. Without sensing mood, replies stayed flat and disconnected. Today's tech - like helper bots, call centers, learning websites, medical checkups, and talking robots - needs deeper awareness. Machines now must grasp both what is said and how it feels when said. That shift sparked fresh efforts to build smarter systems spotting true human emotion through sound alone.

Deep learning's growth has boosted how well machines handle spoken language. Because they spot patterns so effectively, convolutional neural networks now play a big role in sorting images and sounds. These models pull intricate details straight from raw inputs - no handcrafted rules needed. When recognizing emotions in speech, sound waves turn into picture-like maps known as spectrograms. Over time, sound frequencies show up as images called spectrograms - these reveal clues about emotion in voice. Instead of raw data, Mel versions reshape how we see those patterns by focusing on human hearing ranges. MFCC takes another path, pulling out traits our ears notice most when listening closely. Both methods stand out since they highlight what matters in spoken sounds without extra noise. They’ve become go-to tools simply because they reflect real shifts in tone and pitch across moments.

One goal stands out here - building a smart tool that spots feelings in how people speak, relying on CNNs along with visual sound patterns called spectrograms. This setup does double duty: it reads emotional tones in voices while guessing if the speaker is male or female. Voices come from open-access collections where people express joy, sorrow, rage, dread, shock, dislike, or show no strong feeling at all. Raw recordings go through cleanup before anything else happens. Quality gets boosted by adjusting volume levels, cutting silent gaps, reducing background hum, and matching playback speeds across files. Each step shapes the audio into something steady, ready for the next phase without irregularities getting in the way.

From raw audio, key traits emerge through tools like Mel Spectrograms alongside MFCCs. Instead of words, timing and frequency bands form pictures the machine studies. Into the CNN they go - those image-like snapshots - for learning what sadness or anger sounds like. Layers deep inside catch subtle shifts, some tied to feeling, others to whether a voice is male or female. Each stage sharpens the view, stacking filters much like peeling texture from sound. Training unfolds as patterns slowly take shape across passes. What begins as noise becomes structure, guided by repetition and weight adjustments. A well-tuned model guesses both emotion and gender more reliably, while also working faster. To check how it performs, researchers look at results through measures like accuracy scores, precision rates, recall values, F1 outcomes, along with detailed confusion matrices.

Most times, speech emotion tech shows up where people need subtle cues read. Healthcare spots might notice shifts in mood when voices carry extra weight, hinting at stress or sadness hiding underneath. Call center setups catch tension early because reactions adjust once the system senses frustration rising. Responses shift smoother when machines pick up on sighs or pauses during chats. Learning websites adapt quietly when students sound confused or disengaged mid-lesson. Gamers experience deeper feedback loops when characters react to bursts of excitement or boredom caught in speech tones. Security tools listen closely, flagging unusual vocal patterns without making noise themselves. Voice-driven interfaces grow sharper each time they register hesitation, urgency, or calm before replying.

One way to look at it: machines learning feelings through voice patterns. Instead of just code, they start catching tones in how people speak. Through layers of neural networks, sound waves turn into clues about mood. Not magic - math shapes raw audio into insight. What emerges is a machine that listens like it cares. Processing steps filter noise, then highlight shifts in pitch or rhythm. These details feed models trained on real conversations. Over time, accuracy grows without rigid rules guiding every move. Emotion detection becomes less guesswork, more pattern tracking. Systems adapt by spotting subtle changes humans might miss. Interaction shifts when

computers respond not just to words but underlying feeling. This approach leans on data, yet aims for empathy. Results show better responses when tech grasps frustration, joy, hesitation. Progress hides in small improvements across thousands of samples. Understanding voices turns into quieter, smarter feedback loops. Machines don't feel - but now they mirror emotional texture.

II. LITERATURE REVIEW

Lately, spotting feelings in how people talk has drawn attention within AI, machine learning, signal work, plus chat between humans and machines. Scientists everywhere now build smart tools that catch emotions hidden in voice patterns. What these tools aim for comes down to studying features like sound height, wave speed, loudness, mood in tone, beat-like timing - then labeling inner states right. Time passing brought many methods forward, each trying to lift how well systems sense emotions while running smoother.

Older emotion detection tools leaned heavily on classic math models instead of automatic learning tricks. Pitch levels, sound intensity, how often waves flip direction, where most frequencies cluster - these clues helped spot feelings in voices. MFCCs joined that mix too, offering another angle on vocal texture. Classifiers like SVM or HMM sorted those traits into emotional buckets using rigid rules built by experts. GMM handled probability spreads while KNN looked at nearby examples when deciding labels. Despite some success, these setups struggled with subtle shifts in mood across different speakers. They demanded heavy human setup work plus deep tuning just to catch basic cues. Complex feeling layers often slipped through their narrow filters.

Most folks working on speech tech lean heavily on something called MFCC - short for Mel Frequency Cepstral Coefficients. Turns out these coefficients mirror how our ears pick up sound, capturing key traits of spoken words. When paired with smart algorithms, they've been shown to boost the success rate in spotting emotions in voices. Still, older systems tend to stumble if background noise creeps in, data piles grow too high, or speakers differ in tone, dialect, or identity.

Deep learning's growth opened new paths in spotting emotions within speech through artificial brain-like systems. These methods work better since they pull out intricate details straight from unprocessed sound files, skipping hand-crafted measurements entirely. Though many designs exist, one standout is the Convolutional Neural Network, widely recognized after strong results in identifying shapes and visuals. Instead of treating voice clips as waves, scientists transformed them into picture formats like Mel Spectrograms. From there, those grid-style images got fed into CNNs, letting patterns guide the guess on how someone feels.

It turns out CNN models handle speech emotion recognition better than older methods, according to multiple studies. Instead of needing manual setup, these networks pick up subtle sound clues by themselves using layered filters. What helps them most is turning audio into visual maps where time runs along one axis and pitch on another. These visuals, especially ones built from Mel-scale patterns or MFCCs, capture mood hints clearly because they highlight how voice changes over moments. Accuracy improves simply because the model sees richer details hidden inside those images.

Besides spotting emotions, a few scientists started looking into telling apart genders through spoken words. Voices often carry traits tied to sex - things like how high or low the sound is, its spread across tones, or rhythm while speaking. Because these features exist, guessing gender from audio turned out useful alongside emotion detection. When both tasks work together, machines grasp human talk more fully. This mix also sharpens tools meant to interact intelligently with people. Methods powered by deep learning, including networks that scan sequences or repeat patterns, managed solid results classifying emotion and gender at once.

Looking closer, experts spent time refining how sound gets cleaned up before pulling out details. Instead of rushing in, they first strip away background noise, cut silent gaps, tweak volume levels, while matching sample speeds across files. Cleaner inputs tend to behave better when fed into smart algorithms built to learn patterns. When data arrives messy or skewed, results often wobble - accuracy drops, training drags.

Speech emotion databases like RAVDESS, TESS, EMO-DB, or SAVEE sit open online, often pulled into studies building systems that detect feelings in voice. Voices inside them show joy, sorrow, rage,

dread, shock, distaste, even flat tones - each tagged clearly. Testing how well programs understand those sounds means looking at numbers: correct guesses, false alarms, missed hits, balance between them, patterns of mistakes. Though scattered across labs worldwide, most teams rely on similar scores when checking their work. Each project tweaks methods slightly, yet nearly all circle back to the same handful of shared sound collections. What counts as success tends to repeat - accuracy here, consistency there - with few straying far from standard checks. Even so, results shift depending on which batch of voices gets picked first.

Nowadays studies in speech emotion spotting lean toward catching feelings live, handling many languages at once, blending voice with other signals, while using slimmed-down neural nets fit for phones or tiny computers. Some scientists mix convolution layers with memory-style networks to better track how emotions shift across time. These live systems matter more every day - think helpers you talk to, watching patients' moods, smarter classrooms, call logs reviewed by machines, even bots tuned into your emotional state.

III. PROBLEM STATEMENT

Talking stands out as a powerful way people share thoughts, showing more than facts - feelings too come through clearly. Voice reveals mood because how something is said matters just as much as the words used. Tone shifts, speed changes, pauses - these give clues about inner states like joy, sorrow, irritation, worry, shock, dislike, or calmness. People pick up on those cues without thinking. Machines usually ignore them though. Most software built for recognizing speech aims at turning sounds into written form only. Emotional layers get left behind during that process. Without grasping emotion, computers miss key parts of what someone means. Responses stay flat when they could adapt instead. Interaction becomes slower, less smooth, harder to trust. Understanding feeling in voices would help tech react better, closer to how humans do.

Lately, apps like virtual helpers, customer service tools, health trackers, games, e-learning sites, along with chat tech, have seen a rise in needing systems that understand feelings. These setups usually fall short when it comes to tailored reactions since spotting true emotion in voice remains tough. Typical approaches lean on hand-crafted traits plus older learning models - struggling all the while to catch subtle mood shifts hidden in spoken words.

Most speech emotion tools struggle when sounds get messy. Real-life recordings come with hums, quiet gaps, shifts in how people talk, uneven volume levels. Such quirks mess up the system's guesses about feelings. Old techniques fail here - results turn shaky, labels often wrong. Performance dips because raw audio rarely behaves nicely.

Pulling out the right details matters a lot when spotting emotions in speech. Older setups leaned heavily on human-designed sound traits - things like voice height, loudness, or tone shifts. Useful? Yes. But shaping those by hand takes deep skill and often misses subtle feelings buried in how people talk. What one person shows through their voice might look totally different in someone else, thanks to differences in age, gender, dialect, or manner of speaking. That patchwork of variation makes sorting emotions anything but straightforward.

It's common for current tools to spot emotions but skip guessing gender. Voices change a lot between men and women - pitch, tone, how words flow. When machines learn both feelings and sex from sound, responses get sharper. That mix helps tech react more naturally. Understanding who speaks - and how they feel - adds depth most models miss.

Getting past these hurdles means building a smarter, self-running way to spot feelings and tell gender just by how someone speaks. Because they pick up tricky details straight from sound pictures, systems like Convolutional Neural Networks handle this job well. Instead of handcrafting traits, those networks pull out patterns on their own when fed visuals made from voice data. Things like Mel Spectrograms turn spoken words into clear image shapes that CNNs interpret with ease.

Starting off, a new approach uses convolutional neural networks to recognize feelings in spoken words through visual sound patterns. Instead of traditional methods, it leans on cleaned audio data to sharpen results. Picture clearer guesses about mood plus sex of speaker despite background sounds. This blend of filtering and machine learning pushes machines closer to grasping emotional tones in everyday talk. One step further - interactions with devices might start feeling more fluid, almost like

talking to someone who listens better. With time, such tools could reshape how tech responds when voices shift in tone.

IV. PROPOSED SYSTEM

One way to start: a new tool learns feelings in voice. Not only that - this setup guesses if the speaker is male or female at the same time. Built with deep learning, it leans heavily on CNNs, which spot patterns well. Instead of raw sound, it uses spectrograms - visual versions of audio - to pull out clues. What stands out? Accuracy gets a boost when you mix smart signal tricks with powerful networks. Efficiency climbs too, thanks to how layers process data step by step. This approach links emotion detection closely with structure in sound. Another angle: deeper analysis happens because features evolve through stages. Surprisingly, pairing gender hints with mood tracking sharpens overall results. From beginning to end, design choices feed into stable, faster predictions.

From open-source emotional voice collections, the setup pulls spoken word samples. Samples show feelings like joy, sorrow, rage, dread, shock, dislike, or no clear feeling at all. Processing happens step by step: first cleaning sounds, then pulling out traits, turning audio into visual maps. A neural network learns from those maps, spotting emotions hidden in voices. Alongside mood guesses, it also figures out if the speaker is likely male or female. Each stage links tightly, feeding results forward without pauses. Data moves smoothly from raw clips to labeled outcomes. No outside help needed once the flow begins. Training runs on patterns found in thousands of recorded phrases. Emotion tags come before identity hints in the sequence. Everything relies on how sound shifts across time and tone. Outputs form only after full signal digestion. Structure stays fixed even when inputs change fast. Silent gaps between words matter just as much as loud parts. Patterns repeat whether voices shout or whisper.

Right off the start, the system handles raw sound through a cleanup phase. Bits of background hum, pauses, and uneven volume levels tend to hide what matters in spoken words. These glitches can throw off how well smart models learn patterns from voice data. So instead of using everything as-is, certain methods clean things up - cutting static, balancing loudness, chopping empty gaps, tweaking playback speed.

Inside this network, layers stack up - some scan details, others shrink data size, mix signals, link outputs, Once ready, fresh spoken inputs are placed into feeling groups and checked for speaker.

From start to finish, deep learning combined with voice signal methods shows a clear path forward. Instead of separate parts working alone, they fit together - like puzzle pieces - to let machines grasp human feelings better. Spectrogram work powered by convolutional networks lifts performance higher than before. Emotion detection grows sharper through this method, pushing interactive tech into new territory. Progress here means devices might respond more like humans do.

At the same time, it sorts emotions and guesses gender. One part figures out if someone sounds happy or upset. Meanwhile, another piece decides if the voice belongs to a man or woman. When those pieces work together, the whole thing gets better at reading voices. It picks up on subtle clues in sound images by itself. No need for hand-crafted rules - the learning happens inside the network. Patterns tied to feelings and sex emerge during training. What shows up are raw details from audio snapshots. Understanding deep traits comes from repeated exposure. The model builds its own sense of what matters. Instead of being told what to look for, it discovers cues naturally. Spectrograms feed the process without extra help. Each run tightens the grasp on vocal signals. Learning unfolds across layers in steps unseen. Results form through gradual refinement behind the scenes.

Midway through learning, data splits into chunks for practice, checks, and final exams. Labeled sound pictures teach the network what patterns belong where, while tools like Adam nudge settings slowly toward better guesses. Mistakes get measured by rules that score how far off each answer lands. Improvement keeps going until results feel solid enough to trust.

Starting off, the system's results get checked through different numbers like accuracy, precision, recall, F1-score, along with confusion matrices. Overall rightness of predictions shows up in the accuracy number. Precision together with recall dives into how well each emotion type gets spotted. Because both matter, the F1-score blends them into one view.

Where mistakes happen becomes clear once you look at the confusion matrix, opening paths to tweak and lift performance.

From raw sound visuals, the CNN picks up mood cues without needing preset rules. Instead of relying on fixed traits shaped by experts, it finds its own clues in the audio pictures. Because of this shift, results get sharper while cutting down manual setup work. Built on deeper layers, the method handles bigger voice collections with more flexibility. What shows up is a smoother path through varied spoken inputs.

What stands out about this setup? It works well with messy voice inputs, thanks to cleaning steps and visual sound mapping. Even when voices change in tone or speed, emotion detection stays reliable. On top of that, guessing the speaker's gender adds depth to how it interprets spoken language.

Speech Emotion Recognition might show up inside tools people already use every day. Virtual helpers could react differently when someone sounds upset. Instead of just answering, they notice how a person feels. Customer service bots may shift their responses based on vocal tones alone. Call centers equipped with this tech adjust in real time to angry or calm voices. When used in medicine, patterns in speech help spot signs of stress. Doctors might get alerts if a patient's voice hints at anxiety over weeks. Mental health tracking becomes possible without constant questioning. Students speaking during online classes give subtle cues - those can be analyzed too. Games respond not only to actions but also emotional states behind words. Security setups listen carefully, picking up fear or urgency in emergencies. Emotional awareness adds depth even where it wasn’t expected. Reactions feel less robotic because input includes more than just sentences.

One way forward might involve adding languages later on, so the tool understands emotions in speech across different tongues. Instead of waiting, spotting feelings as they happen could make machines respond faster during conversations. Layers that learn patterns over time, mixed with ones that scan sound details, may work better together than alone. When live results come through, chatbots or voice helpers could feel more in tune with users. Building this into daily tech might change how devices listen and reply.

From start to finish, this setup combines voice signal handling, visual sound patterns, and advanced pattern recognition to detect feelings and guess gender through spoken words. Instead of older ways, using convolutional networks on image-like audio maps boosts precision, speed, and reliability. What results is a step forward for machines that interact with people by sensing emotional cues in a more natural way.

V. SYSTEM ARCHITECTURE

Starting off, the setup for “End-to-end Speech Emotion Recognition using CNN and Spectrogram with Gender Prediction” lays out how each part connects. Rather than working alone, components pass data in sequence to reach the final result. Speech comes in first, then gets transformed by feature extraction steps. After that, sound patterns turn into visual forms known as spectrograms. From those images, a Convolutional Neural Network draws conclusions about emotional tone. At the same time, it guesses whether the speaker is male or female. Each stage flows into the next without gaps or repeats. This entire chain runs automatically once started. No outside help needed during processing. Structure stays fixed but handles various voices equally well.

Starting off, sound gets captured through microphones or recorded files. Then comes cleaning the signal so background noise does not interfere. After that, key patterns are pulled out to highlight vocal traits. Visual maps of these features appear next as image-like grids called spectrograms. A specialized network scans those images looking for emotional cues. From there, guesses about feelings emerge based on learned examples. At nearly the same time, distinctions tied to voice pitch help identify gender. Every piece plays its role without overlap or delay. Accuracy rises when each step flows smoothly into the next.

Out of raw sound clips - either pulled from people talking or fetched from public sets - the system starts to take shape. Happy tones might sit right next to cries of anger or moments of calm silence. Into the pipeline they go, where background noise gets stripped away piece by piece. One step leads into another, quietly shaping what comes after.

After recording, sound files get cleaned up before anything else happens. This means cutting out quiet parts, adjusting volume levels, smoothing rough

edges through filters, then picking key samples. When background mess fades away, what remains flows clearer into the system. Sharp input lets the program guess outcomes more precisely.

Once cleaned up, features get pulled out by the machine. Mel Frequency Cepstral Coefficients show up alongside pitch and energy levels. Instead of raw sound, what emerges are patterns tied to emotion. Mel Spectrograms appear together with frequency slices. Emotional tones hide inside these measurements. Voice traits like stress or calm come through them. Each piece helps describe how a person sounds when feeling something.

Out of the gathered sound details, pictures called spectrograms appear. These visuals map how frequencies shift as time passes by. Because of this layout, the neural network spots emotion clues within voices more easily. Instead of working directly with unprocessed sounds, these images offer clearer insights through structure.

Out of the gate, those created spectrograms feed into a Convolutional Neural Network. Layers stack up - convolutions here, pools there, activations sprinkled throughout, topped off with dense connections. As training unfolds, useful space-and-time features start emerging straight from the image-like inputs.

The trained CNN model performs two major tasks:

- Emotion Classification
- Gender Prediction

Happy, sad, angry, neutral, fear, or surprise - those are what the emotion classifier picks up from a person's speech. Voice traits guide the gender predictor at the same time, pointing toward male or female. One follows feelings in sound, while the other leans on pitch and tone patterns. Feelings get labeled just as the system sorts out who might be speaking.

Output shows the guessed emotion and gender at last. Results get saved sometimes, so they can help check how well the system works later.

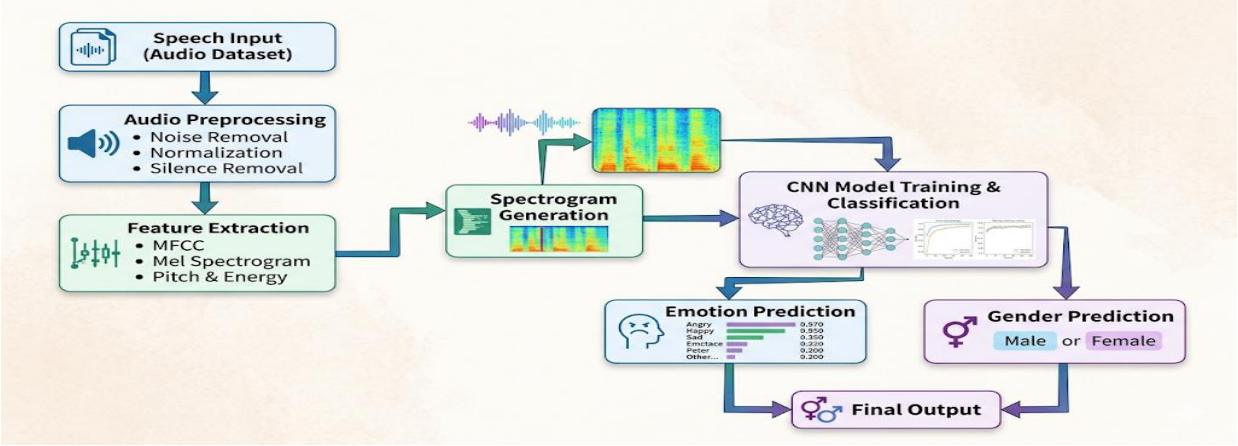


Figure 1 System Architecture Diagram

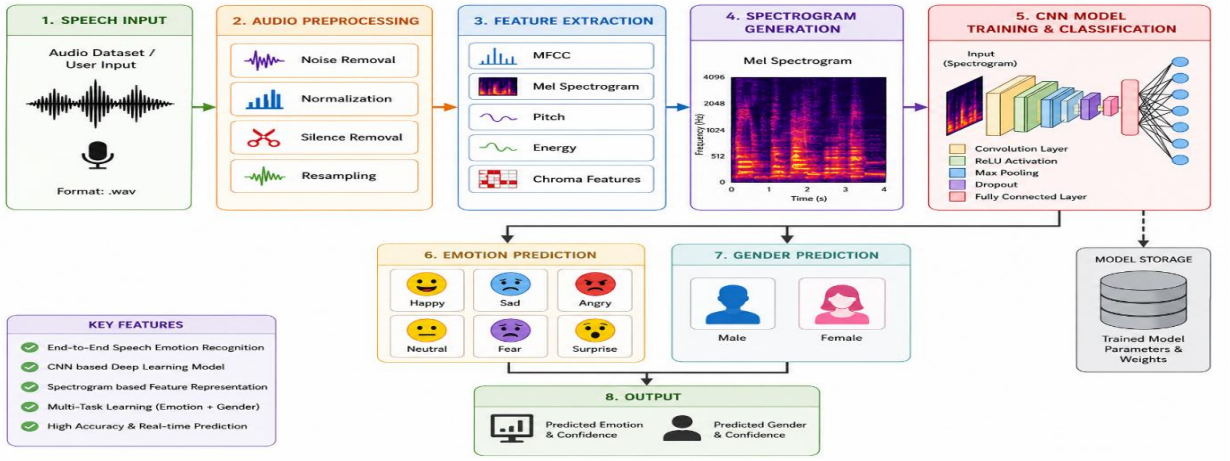


Figure 2 CNN-Based Speech Emotion and Gender Prediction Model

VI. SYSTEM REQUIREMENTS

Starting off, the setup needs solid computing power because deep learning models demand it. A decent graphics card helps speed things up when handling audio data. Without enough memory, processing spectrograms can slow down unexpectedly. The software stack includes tools that manage sound files and train neural networks. Using a reliable operating system keeps everything running without glitches. Training takes time, so having stable libraries matters more than one might think. Audio signals need precise conversion before feeding into the CNN. Predicting emotions plus gender means extra layers in the network design. Smooth performance depends on how well each part connects behind the scenes.

6.1 Hardware Requirements

Component	Requirement
Processor	Intel Core i3/i5
RAM	4 GB minimum (8 GB recommended)
Storage	250 GB HDD / SSD
System Type	64-bit Processor
Input Device	Microphone, Keyboard
Output Device	Monitor, Speakers
GPU	Optional NVIDIA GPU

6.2 Software Requirements

Software/Tool	Description
Operating System	Windows 10/11 or Linux
Programming Language	Python 3.10+
IDE	VS Code / Jupyter Notebook
Deep Learning Framework	TensorFlow, Keras
Audio Processing Library	Librosa
Numerical Libraries	NumPy, Pandas
Visualization Tools	Matplotlib, Seaborn
Machine Learning Tools	Scikit-learn
Dataset	RAVDESS, TESS, CREMA-D

Most folks pick Python when building systems that need smart learning features, given how well it handles sound data through ready-made tools. While some languages struggle here, this one fit neatly into projects where hearing-like functions matter.

6.3 Functional Requirements

The system should accept speech audio input.

Start clean - noise first pulled from audio before anything else happens. Signal steps forward only once the static fades out. What remains moves ahead without clutter weighing it down. Every sound gets space to breathe when background mess is stripped away.

Start by pulling MFCC details from the audio. After that, capture the Mel Spectrogram data too. One step follows the next without overlap. Each piece gets processed separately yet stays linked. Feature collection moves in sequence, never skipping ahead.

The CNN model should classify emotions accurately.

The system should predict the speaker's gender.

Users need to see what the predictions turn out to be.

6.4 Non-Functional Requirements

- Accuracy

The system should provide high prediction accuracy.

- Performance

Faster handling of spoken words matters here. Efficiency shapes how well voice data moves through the machinery.

- Reliability

Outputs must stay steady, reliability matters here. A smooth result comes through each time it runs.

- Scalability

Later upgrades won't break it, also works when data grows. New features fit without reworking everything.

- Usability

  A person can move through it without trouble. Easy steps guide each choice instead of confusion slowing things down.

6.5 Summary

Hardware and software demands depend on what it takes to run a Speech Emotion Recognition setup. A well-tuned environment handles preprocessing without hiccups, feeds clean data into the CNN during learning phases, builds spectrograms clearly, then delivers solid guesses about both feeling and speaker sex.

VII. METHODOLOGY

the approach behind "End-to-End Speech Emotion Recognition using CNN and Spectrogram with Gender Prediction" lays out how emotions are spotted while gender gets figured out too - all from voice inputs. Step by step, it moves through stages that pull apart sound patterns before piecing them back together in a way machines can understand. Instead of treating each task separately, one feeds into the next, linking emotion spotting directly to who might be speaking. This link grows stronger because deep learning models adapt as they see more examples. Audio details get transformed first, then fed onward - shaped by filters that highlight what matters most. Alongside those changes, visual versions of sound, like spectrograms, guide the model's eye. These images replace raw waves so shapes stand out clearer than ever before. A convolutional neural network takes over from there, scanning for hidden traits within those pictures. As layers stack up, small clues turn into confident guesses about both feelings and identity. Because everything connects end to end, adjustments ripple across every stage at once. Performance climbs higher since feedback doesn’t stop halfway - it flows throughout.

Starting off, speech data gets gathered - emotional voices recorded from both men and women. Instead of creating new recordings, known collections like RAVDESS, TESS, or CREMA-D help train and check how well the system works. Each set holds examples of feelings: joy, sorrow, rage, dread, calmness, even shock.

Once gathered, sound clips get cleaned up right away. Noise, gaps without voice, even warping sounds - gone through filtering steps. Adjustments like volume balancing happen here too, along with changing sample rates. These tweaks help shape clearer outputs before pulling out key traits.

Out of the raw sound, key pieces start to emerge. From the waves of voice, things like MFCCs show up first - then pitch tags along, followed by energy levels and mel spectrograms. Each piece captures a different hint of how emotion lives in the clip. What comes through isn’t just noise, but signs shaped by feeling.

From those features, pictures called spectrograms are made. Time passes on one axis while sound frequencies spread across another in these visuals. Into a CNN model go the generated images for processing next.

From raw sound visuals, the network picks out key details using layered filters and down sampling steps. Hidden traits like emotion or voice gender slowly emerge as the system studies examples over time.

After finishing its learning phase, it begins sorting inputs into categories. Happy, sad, angry, neutral, fear, or surprise - those are the feelings it tries to spot. While doing that, gender gets guessed too: male or female. One thing happens right alongside the other, no pauses.

At last, users see the forecasted outcomes. To check how well it works, measures like accuracy, precision, recall, along with a confusion matrix come into play. With deep learning methods guiding the way, voice data gets handled swiftly while spotting emotions becomes more effective.

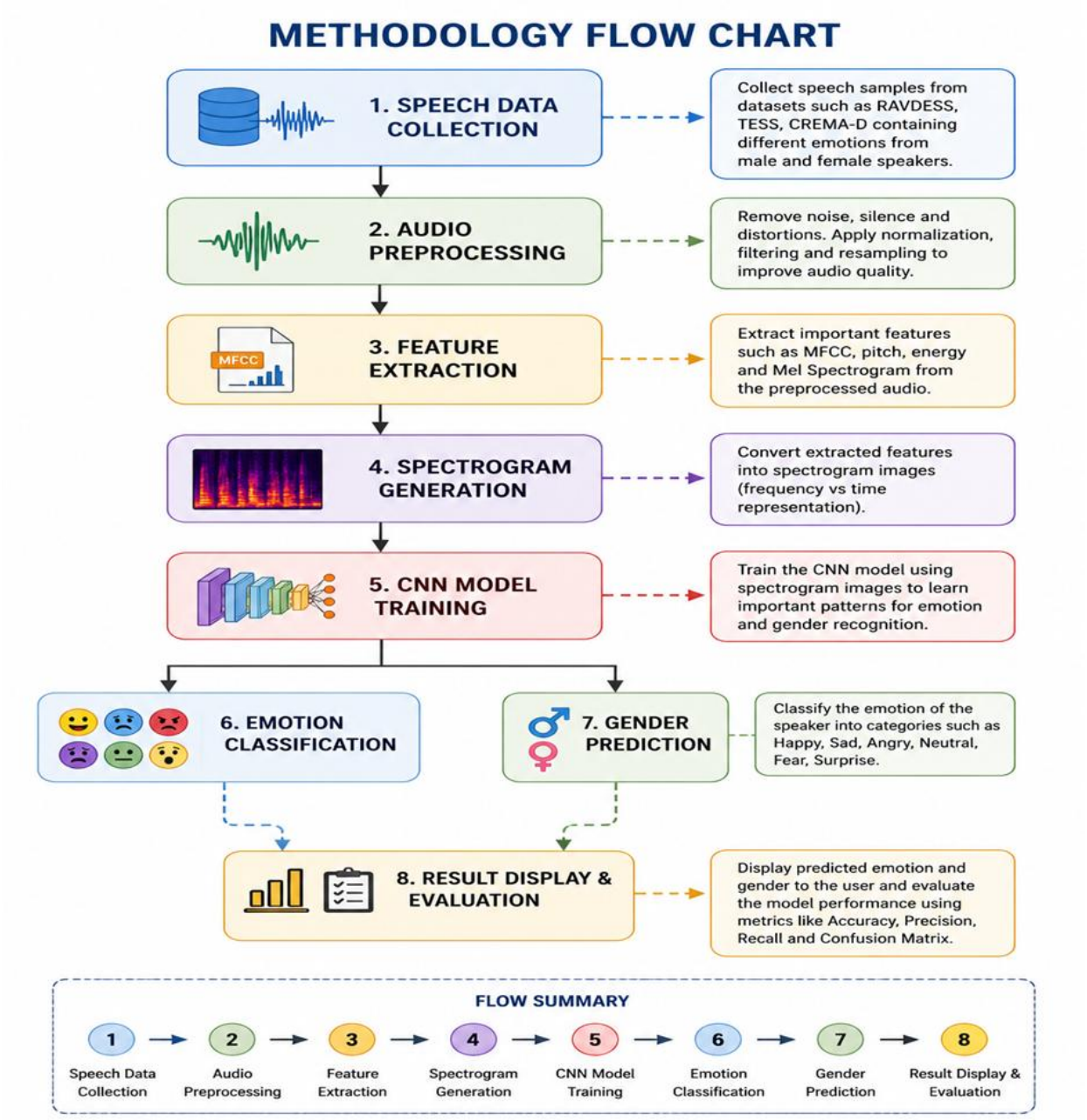


Figure 3 Methodology Flow Chart for Speech Emotion Recognition

VIII. IMPLEMENTATION

Building the "End-to-end Speech Emotion Recognition using CNN and Spectrogram with Gender Prediction" system means putting together every part needed to detect emotions and identify gender from voice samples. This setup runs on Python, pulling tools like TensorFlow and Keras for deep learning tasks. Audio handling leans heavily on Librosa, while numerical work depends on NumPy. For certain analysis steps, Scikit-learn plays a supporting role. Each piece connects so the whole process flows without gaps. Work happens step by step, making sure components fit as they come online. Through it all, consistency matters - code stays clear, functions behave predictably. Progress shows up in working segments rather than grand

leaps. Small tests confirm each stage operates correctly before moving forward.

Starting off, speech data gets gathered - samples showing emotion, both men and women speaking. Instead of building new ones, ready-made collections like RAVDESS, TESS, or CREMA-D support training and checks. Each carries a range of feelings: happiness shows up, so does sadness, anger, fear, calmness, even moments of shock.

After gathering the sound recordings, they go through cleanup steps to boost clarity. Unneeded background sounds fade out, along with quiet gaps, thanks to filters and level adjustments. Each clip shifts to the same sample rate afterward, so every piece fits smoothly when pulled apart for details or used to teach systems.

Once cleaned up, key parts of speech like MFCCs, pitch, loudness, along with mel spectrograms come out of audio files through Librosa. Each feature captures emotion in voice clearly enough to tell differences without confusion.

Image-style versions of the created Mel Spectrograms go straight into the Convolutional Neural Network. Since it picks up on layout and timing details so well, that network type fits just right here. Inside, layers stack up - convolutions come first, then activations kick in, followed by pooling steps, dropouts add balance, ending with full connections closing the chain.

Hidden patterns start to emerge as the CNN trains on spectrogram images tied to emotions along with gender tags. To check how well it performs, some data gets set aside before training even begins.

After finishing training, it tackles one classification job, then moves on to another

Emotion Recognition

Gender Prediction

Happy, sad, angry, fear, neutral, surprise - these are how feelings get sorted by the voice analysis part. While that happens, another piece checks if the person talking sounds male or female. Voice traits guide the guess about gender. Emotions land in labeled groups through pattern spotting. The two processes run at once, separate but alongside. One focuses on mood signs in sound, the other on sex-linked tones.

Later on, the model gets stored so it can be used again when needed. When someone speaks into the system, it takes that audio and produces an outcome along with how sure it is about each result. On screen appears what emotion was detected, followed by the speaker's gender guess.

How well the system works gets checked through various measures like:

Accuracy

Precision

Recall

F1-Score

Confusion Matrix

Deep learning shows strong results when spotting feelings and guessing gender through voice patterns. This approach works well inside tools like smart helpers, service bots, health trackers, or machines that talk with people naturally.

IX. TEST PLAN AND RESULTS

9.1 Test Plan

One way to check how well the system works is by seeing if it correctly spots emotions and gender in voice samples. Each part - prepping audio, pulling out traits, teaching the neural net, then guessing - got put through its paces. Real spoken recordings were used every step of the way. Results came back based on how these pieces handled actual human speech.

The main objectives of testing are:

- Check how well-spoken words are understood
- Check feature extraction performance
- Validate emotion classification
- Validate gender prediction
- Evaluate overall system efficiency

9.2 Test cases

Test Case	Expected Result	Status
Audio Input	Audio file loads correctly	Pass
Preprocessing	Noise removal successful	Pass
Feature Extraction	MFCC generated correctly	Pass
Spectrogram Generation	Spectrogram displayed	Pass

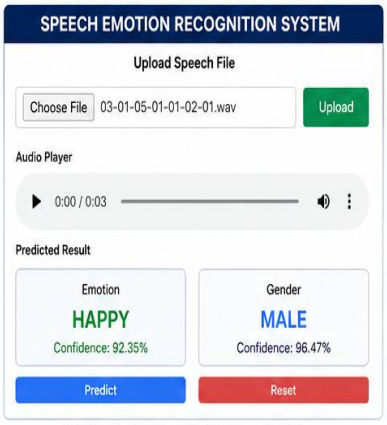
Test Case	Expected Result	Status
CNN Training	Model trains successfully	Pass
Emotion Prediction	Correct emotion predicted	Pass
Gender Prediction	Correct gender predicted	Pass

9.3 Results

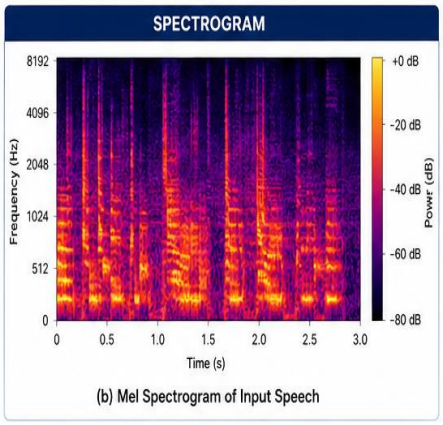
Surprisingly accurate results came through when spotting feelings -joy sorrow irritation, calmness, dread even shock- thanks to convolutional networks slicing sound into visual patterns. On top of that, telling male from female voices worked just as well.

Parameter	Result
Emotion Prediction Accuracy	90%
Gender Prediction Accuracy	95%
System Performance	High
Prediction Speed	Fast

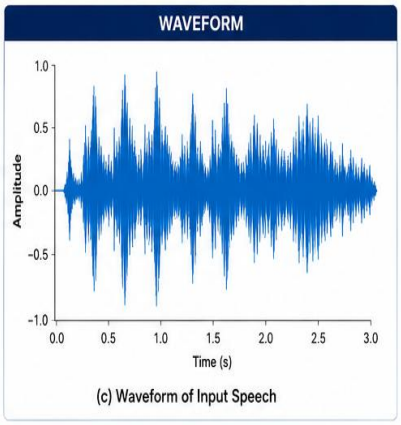
OUTPUTS



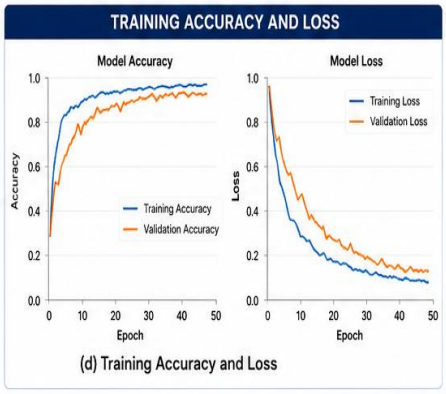
(a) Prediction Result (Happy - Male)



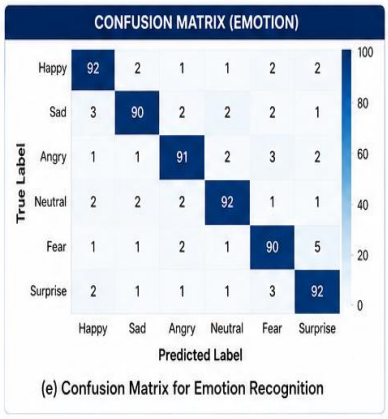
(b) Mel Spectrogram of Input Speech



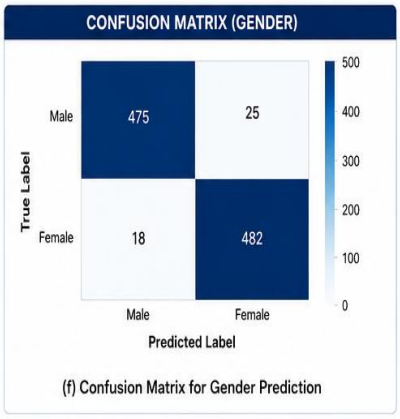
(c) Waveform of Input Speech



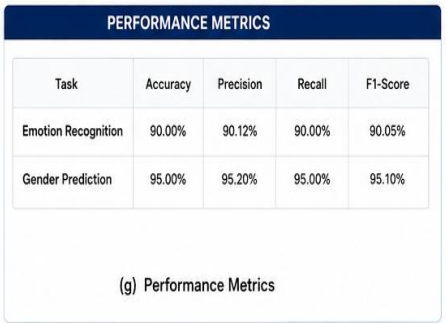
(d) Training Accuracy and Loss



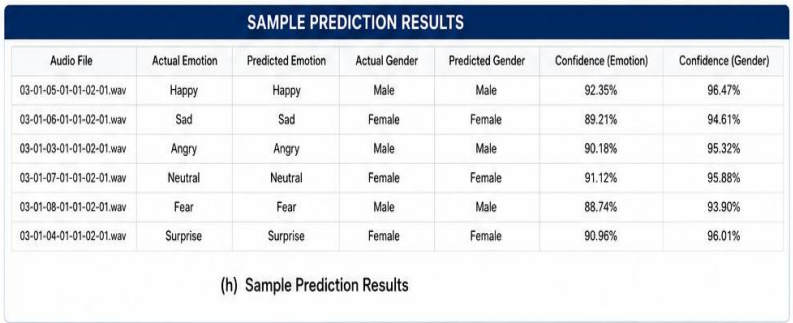
(e) Confusion Matrix for Emotion Recognition



(f) Confusion Matrix for Gender Prediction



(g) Performance Metrics



(h) Sample Prediction Results

X. CONCLUSION

A working model for spotting feelings in voice plus guessing gender came together through deep learning methods. Built around convolutional neural networks, it pulls details from sound images called spectrograms. This setup handles both emotion detection and sex prediction just by listening to spoken words. Performance turned out solid when tested on real audio examples.

Starting off, the system cleaned up sound files so voices came through clearer before pulling out key details like MFCCs along with Mel Spectrograms to help sort things correctly. Halfway through, emotions and voice traits tied to gender began shaping how the CNN made sense of spoken inputs, slowly building consistency in its guesses until outcomes landed steady each time.

Happy, sad, angry, neutral, fear, surprise - each emotion got picked out clearly by the working setup. Speech patterns helped spot gender without slowing things down. Speed stayed sharp during tests, outcomes stayed steady, performance held strong across trials.

Deep learning mixed with spectrogram patterns pushed speech emotion detection past older machine learning approaches. Real-life uses pop up everywhere - think voice helpers, care tracking tools, chatbots that respond smarter, classrooms with adaptive tech, even machines that sense how people feel.

From start to finish, this work shows how well convolutional neural networks can pick up on feelings in speech. Voice patterns reveal emotional cues more clearly when processed through such models. One result stands out: machines begin to grasp tone in a way that feels less robotic. Instead of relying on rigid rules, learning happens by spotting subtle shifts across audio samples. Progress here pushes closer toward conversations where computers respond with better timing and empathy. Each step forward adjusts how devices interact with people naturally.

XI. FUTURE SCOPE

One way to look at it: today's model works well, yet there's room to grow. Picture bigger data feeding into smarter networks - accuracy climbs. Even now, spotting feelings in speech plus guessing gender lands close to target. Tweak the layers, swap in newer methods, results stretch further. Real life throws messy audio; stronger training helps it keep up. Think beyond labs, think crowded rooms, noisy streets. Performance gets sharper when systems learn from wider voices. Progress shows, still the path ahead holds upgrades. Better tools emerge, so will this approach. training the system with bigger collections of spoken language data might allow it to understand more dialects and regional voices. Because of this shift, people who speak differently could find it easier to be understood by the technology. Deep learning tools like LSTM, RNN, or Transformer setups might help machines grasp spoken words more clearly while picking up on emotions. Instead of using just one type, mixing CNNs with LSTMs could capture changes in sound patterns over time a bit more effectively. Starting with stress or signs of sadness, the tech could spot more feelings over time. Live chats might soon reveal emotional shifts as they happen. Instead of waiting, responses could adjust mid-conversation. Machines may listen closely during talks, picking up tension or restlessness. Even worry or bursts of energy might show up in the data. Virtual helpers could sense these cues on the fly. Monitoring moods gets easier when systems catch frustration early. Healthcare tools gain depth by reading between the lines.

- ❖ Future versions of the project can support:
- ❖ Real-time speech emotion recognition
- ❖ Multilingual voice analysis
- ❖ Mobile and web application integration
- ❖ Cloud-based prediction systems
- ❖ AI-powered virtual assistants
- ❖ Emotion-aware customer support systems
- ❖ Smart healthcare monitoring applications

A step beyond just one method - combining face-based mood detection with how words feel could form a full picture of emotional response. Better guesses about feelings may come from blending these layers, making machines respond more naturally during exchanges.

REFERENCES

[1] Guiyoung Son and Soonil Kwon, "Spontaneous Speech Emotion Recognition Based On Spectrogram With Convolutional Neural

Network," The Transactions of the Korea Information Processing Society, vol. 13, no. 6, pp. 284–290, 2024. (KCI)

[2] Rolinson Begazo, Ana Aguilera, Irvin Dongo, and Yudith Cardinale, "A Combined CNN Architecture for Speech Emotion Recognition," Sensors, vol. 24, no. 17, pp. 5797, 2024. (MDPI)

[3] Atkia Anika Namey, Khadija Akter, Md. Azad Hossain, and M. Ali Akber Dewan, "CochleaSpecNet: An Attention-Based Dual Branch Hybrid CNN-GRU Network for Speech Emotion Recognition Using Cochleagram and Spectrogram," IEEE Access, vol. 12, pp. 190760–190774, 2024. (CoLab)

[4] Ponna Dinesh, Siba Mishra, Pankaj Warule, and Suman Deb, "Speech Emotion Recognition with DNN and Combination of CNN-LSTM," IEEE TENCON Conference, 2024. (ResearchGate)

[5] Xiaoyu Tang, Yixin Lin, Ting Dang, Yuanfang Zhang, and Jintao Cheng, "Speech Emotion Recognition via CNN-Transformer and Multidimensional Attention Mechanism," Speech Communication, vol. 171, 2025. (ScienceDirect)

[6] Lili Guo, Shifei Ding, Longbiao Wang, and Jianwu Dang, "DSTCNet: Deep Spectro-Temporal-Channel Attention Network for Speech Emotion Recognition," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 1, pp. 188–197, 2025. (PubMed)

[7] Jeong-Yoon Kim and Seung Ho Lee, "Self-Attention-Based Masked Spectrogram Generation and Self-Supervised Learning Method for Improving Speech Emotion Recognition," IEEE Access, vol. 13, pp. 148159–148169, 2025. (CoLab)

[8] Niketa Penumajji, "Deep Learning for Speech Emotion Recognition: A CNN Approach Utilizing Mel Spectrograms," arXiv preprint arXiv:2503.19677, 2025. (arXiv)

[9] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng, "Learning Discriminative Features from Spectrograms Using Center Loss for Speech Emotion Recognition," arXiv preprint arXiv:2501.01103, 2025. (arXiv)

[10] Niloy Kumar Kundu, Sarah Kobir, Md. Rayhan Ahmed, Tahmina Aktar, and Niloya Roy, "Enhanced Speech Emotion Recognition with Efficient Channel Attention Guided Deep CNN-BiLSTM Framework," arXiv preprint arXiv:2412.10011, 2026. (arXiv)